



ИНСТИТУТ ЗА МАТЕМАТИКУ И ИНФОРМАТИКУ
ПРИРОДНО-МАТЕМАТИЧКИ ФАКУЛТЕТ
УНИВЕРЗИТЕТ У КРАГУЈЕВЦУ

МАСТЕР РАД

**Примена алгоритма моделовања тема из
текстуалних садржаја за проналажења
одговора на постављено питање**

Студент:
Јелица Васиљевић

Професор:
др Милош Ивановић

Август 2015.

Садржај

1	Увод	3
1.1	Опис проблема	3
1.2	Опис решења	3
1.3	Терминологија	3
2	Алгоритам моделовања тема	5
2.1	<i>Latent Dirichlet allocation</i>	6
2.1.1	Генеративни процес	7
2.1.2	Како се откривају теме - упрости́ен пример	8
2.1.3	Графички пример моделовања тема	12
3	Математичка позадина	13
3.1	Теорија вероватноће	13
3.1.1	Основни појмови	13
3.1.2	Условна вероватноћа	14
3.1.3	Случајне променљиве	16
3.1.3.1	Дискретне случајне променљиве	17
3.1.3.2	Непрекидне случајне променљиве	19
3.1.3.3	Вишедимензионалне случајне променљиве	20
3.2	Важније расподеле	24
3.2.1	Биномна и полиномна(енг. multivariate) расподела	24
3.2.2	Дирихлеова расподела	25
3.3	Гибсово узорковање	27
3.3.1	Марковљеви ланци	27
3.4	Како ради ТМ алгоритам	30
3.4.1	Имплементација - псеудокод	34
4	Историја ТМ	35
5	Опис приступа решавању проблема	36
5.1	Припрема података - предпроцесирање	36
5.1.1	Уклањање HTML ознака и неалфанумеричких карактера	36
5.1.2	Конвертовање свих слова текста у „мала слова“ енгл. lowercase	37
5.1.3	Издавање атомских елемената докумената - токена, енгл. tokenization	37
5.1.4	Изабацивање често коришћених речи енгл. stop words	37
5.1.5	Додавање синонима	38
5.1.6	Склањање наставака речи - енгл. stemming	38
5.1.7	Свођење на коренску реч - енгл. lemmitization	38
5.2	Развој решења	38
5.2.1	Општи преглед пакета Mallet	38
5.2.2	Pipe класа Mallet-y	38
5.2.3	ParallelTopicModel класа Mallet-y	38

5.3	Тестирање решења	38
5.3.1	Општи приступ тестирању	38
5.3.2	Начин тестирања у конкретном проблему	38
6	Решење проблема применом методе бројања речи	39
6.1	Опис решења методом бројања речи	39
7	Резултати	40
7.0.1	Утицај стеминга на резултат	40
7.0.2	Утицај додавања синонима на резултат	40
7.0.3	Укупни резултати са синонимима и стемингом	40
8	Коришћене скраћенице	41

Глава 1

Увод

1.1 Опис проблема

Проналажење одговора на постављено питање претстваља свакодневни проблем. Сваким Гугл упитом, покреће се низ алгоритама коју покушавају да одгонетну шта упит заправо представља, које би странице биле релевантне и у ком редоследу. Мера "доброг" одговора на постављено питање знатно може да варира у зависности од сврхе система. Интуитивно, тематика одговора и питања може да послужи као добар критеријум одабира квалитетних одговора. Задатак овог рада је испитивање да ли заиста тематика може да помогне у проналажењу адекватног одговора и у коликој мери.

1.2 Опис решења

Циљ овог рада је израда прототипа програма који би коришћењем алгоритама за моделовање тема из базе потенцијалних одговора проналазио најбољи одговор за задато питање. Дакле, програм не треба да "осмисли" одговор на задато питање већ само да "препозна" који од могућих одговора највише одговара постављеном питању.

Примена оваквог решења могла би да буде значајна у различитим областима од комерцијалних до научних. На пример, омогућило би се ефикасно аутоматско одговарање на често постављена питања која могу имати различиту формулацију или ефикасно проналажење адекватних научних радова.

У раду су обрађивани текстови на енглеском језику али због природе модела, развијени програм и резултате могуће је применити и на било који други језик. Поред основног текста питања и одговора, у раду је испитан и утицај додавања синонима на проналажење одговора као и утицај свођења речи на коренске (склањање глаголских и именских наставака, енг. *stemming*).

Као компаративни модел коришћен је приступ проналажења одговора на основу броја заједничких речи (енг. *WordCount*).

1.3 Терминологија

Општи преглед значења термина који су коришћени у раду као што су реч, речник, вокабулар, корипус итд.

- Тема : скуп речи које је најбоље карактеришу. На пример тема рачунарство би представљала скуп речи : алгоритам, процесор, кодирање, израчунавање, меморија, рачунар, бит, бајт, лаптоп итд. Важно је приметити да неке речи могу припадати у више тема, као што је нпр. реч израчунавање која може припадати и области математика.

- Речник или корпус - скуп свих различитих речи које се јављају у неком скупу докумената
-

Глава 2

Алгоритам моделовања тема

Сваким даном повећава се количина доступних дигиталних информација. Парадокс данашњег времена је да се упрокс великој количини података из различитих области све теже долази до података који су од интереса. Дакле, потребно је пронаћи алат којим би се велике количине података организовале а самим тим боље разумеле и лакше претреживале.

Тренутно, најпопуларнији начин претреге је према кључним речима. Кључне речи се предају неком систему за претрагу а као резултат добијамо скуп докумената који су повезани са њима. Иако овакав систем ради јако добро и са великом поузданошћу, постоје и другачији приступи

Најчешће питање које се поставља захтева одговр из неколико, ужих или ширих области. Међутим, кључне речи које се предају као критеријум претраге могу карактерисати и области које нису од интереса. На пример, Гугл претрга за кључне речи "ген, еволуција" ће у највећем броју случајева садржати веб странице посвећене биологији или сродним областима. Међутим, поменуте речи припадају и области рачунарства (генетски алгоритми) али ти резултати ће знатно слабије бити заступљени у односу на резултате везане за биологију. Тренутно се оваква врста проблема може решити додавањем још неке кључне речи која припада захтеваној области (на пример за задате речи : алгоритам, кодирање, програм итд.) међутим поставља се питање избора адекватних додатних речи, њихове заступљености у захтеваној области као и релевантности у односу на друге кључне речи (на примеру гугл претраге, често се дешава да механизам за претраживање испусти неку од речи и прикаже само резултат за остатак упита)

Један од начина да се поменути проблем реши је и претрага по тематици или теми докумената. Дакле, уместо претраге по кључним речима, најпре се пронађе област у којој се врши претраживање а затим се претражују документи који припадају тој области. На тај начин елиминишу се документи који би се у резултату појавили на основу предатих кључних речи али који нису повезани са тематиком у којој се тражи одговор.

На пример, нека је потребно пронаћи све чланке листа "Политика" у којима се говори о спортским успесима наше кошаркашке репрезентације. Сви чланци овог листа могу се поделити у неколико категорија : политика, хроника, спорт, време итд. Пошто је од интереса категорија спорта, у обзир претраге долазе једино спортски чланци са тематиком кошарке. На овај начин могуће је пратити како се мењао успех кошаркашке репрезентације са временом, колико се обаћала пажња на такве догађаје у ком временском периоду итд.

Описани вид претраге подразумевао би да се сваки лист "Политике" најпре прочита а затим раздвоји по категоријама и "разуме" шта која категорија представља како би се добили тражени подаци. Због велике количине података, овако нешто је немогуће урадити без помоћи рачунара. Алгоритми моделовања тема представљају први корак ка решавању оваквих и сродних проблема.

Циљ алгоритама за моделовање тема је "отривање" тема присутних у некој колекцији докумената. У суштини, алгоритми за моделовање су статистичке методе које на основу

анализе свих речи у документу откривају које су то теме заступљене у том документу. За рад ове методе није потребно никакво претходно означавање докумената, теме докумената зависе једино од речи које се јављају у тексту.

Моделовање тема омогућава организацију велике количине података на нивоу који је тешко дохватљив људским могућностима.

Надаље ће се паралелно употребљавати две еквивалентне ознаке - моделовање тема или ТМ (скраћеница од енг. *topic model*)

2.1 *Latent Dirichlet allocation*

Latent Dirichlet allocation, надаље LDA, је најједноставнији приступ проблему моделовања тема [2] а његова примена била је и предмет овог рада.

Добро је познат роман Бранка Ћопића "Орлови рано лете". Уколико би неко ко није прочитао ови књигу желео да зна "о чему се ради"у њој, највероватније би добио одговор да је у питању књига која се бави доживљајима групе дечака на почетку Другог светског рата. Иако је то најшири оквир романа,у њему су присутне и теме о љубави, дружењу, пријатељству, рату, пустоловинама итд. Дакле, роман ,опште гледано, обухвата више тема, али се са неколико њих интензивно бави.

Уколико сада посматрамо овај рад као један пример документа, у њему се највише "говори"о моделовању тема и њиховој примени, али исто тако, само у мањој мери, и о књижевности (пример књиге "Орлови рано лете "). Дакле, тешко је пронаћи било какав документ који се бави само једном темом. Чак се и у радовима који се баве неким уским научним областима, могу пронаћи делови који се могу сврстати и у неке друге научне области, било из исте, сродне или потпуно различите научне гране.

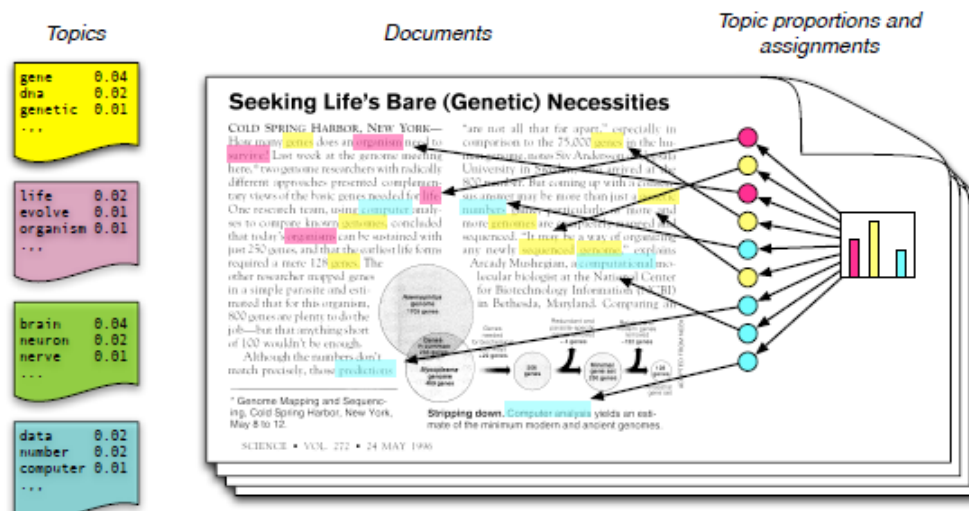
Основна претпоставка LDA модела је управо да се сваки документ може сврстати у више области, тј. да се бави са неколико тема.

Поставља се питање како то, да након читања неког документа, знамо да су у документу присутне различите теме и које су то теме. Уколико неки рад који се бави тематиком машинском учења прочита човек који се не бави рачунарством, највероватније је да би он као теме рада издвојио рачунарство и математику. Међутим, ако би исти рад прочитао неко ко се бави рачунарством, тада би он могао као потенцијалне теме да изведе једну или више области машинског учења, алгоритме, базе података, статистику, теорију бројева итд. Дакле, издвојене теме зависе од "знања"онога ко чита рад. Уопштено, у оба случаја теме су издвојена на основу речи које се најчешће срећу у областима које појединац познаје. Дакле, може се уопштити да тема представља **скуп карактеристичних речи**.

На следећој слици представљен је чланак "*Seeking Life's Bare(Genetic) Necessities*" који "говори о"употреби анализе података за одређивање броја гена који организам треба да има да би преживео (у еволутином смислу).Може се уочити да су три најзаступљеније области у овом тексту - анализа података, еволутивна биологија и генетика. На њој су "ручно"означене неке речи које припадају овим областима. Речи које се могу сврстати у област *анализе података* су означене плавом бојом, речи које припадају *генетици* су означене жутом бојом док су речи које се односе на *еволутивну биологију* означене розом бојом. Уколико би се ова процедура применила на сваку реч текста, јасно би се уочило колико је која тема заступљена у овом тексту. Математички, "*присуство*"теме у тексту се означава односом броја "обојених"речи у једну боју и укупног броја речи у тексту.

Наравно, постоје речи које се могу сврстати у више од једне теме. Такве речи ће бити обојене са две или више боја,али због прегледности слике, такви случајеви су изостављени.

На левој стани слике дате су неке теме (енг. *Topics*) које са одређеним вероватноћама садрже неке речи. На основу припадности речи темама, извршава се описани процес означавања ("бојења") речи да би се на крају добили удели тема у тексту (десна страна



Слика 2.1: Пример чланка, преузето са [2]

слике, енг. *Topics*) Важно је приметити да у тексту постоји доста речи које не одређују ни једну конкретну тему и које се са готово истим, малим вероватноћама могу сврстати у сваку од њих. Такве речи су нпр. везници, личне заменице ,прилошке одредбе итд. Оне се једним именом називају енг. *stop words*. Како присуство таквих речи не утиче на тематику документа, то их при математичкој анализи текста треба занемарити.

LDA је статистички модел који формално описује описани процес означавања речи. Да би се у потпуности разумело како LDA ради, потребно је упознати се са *генеративним процесом* - процесом којим се креирају документи са становишта LDA-а.

2.1.1 Генеративни процес

Нека је дат неки скуп речи најчешће коришћених у научним областима математике, физике, хемије, музике, рачунарства и биологије. Нека је, даље, потребно креирати документ који има одређен број речи из датог скупа речи али тако да он највише "говори о"математици и музици, али поред ових тема, у мањој мери, "говори о"физици и програмирању, док се остале теме занемарљиво мало помињу. Како је дат фиксан скуп речи - речник (вокабулар), могуће је свакој речи придружити *вероватноћу* припадања свакој од тема. Тако ће на пример реч *интеграл* имати велике вероватноће припадања темама математика и физика, мање вероватноће у темама хемија и рачунарство док ће се са јако малим вероватноћама јављати у осталим темама. Дакле, једна реч припада свакој од датих тема али са различитим вероватноћама.

Генерисање траженог документа могуће је извести следећом процедуром :

1. Одабрати расподелу тема у документу - прецизирати која тема се са којим уделом појављује у документу. У конкретном примеру, расподела би могла бити : математика 30%, музика 30%, физика 20%, програмирање 15%, биологија 2% и хемија 3%.
2. Докле год није достигнут тражени број речи
 - (а) Изабрати тему из дистрибуције која је одабрана у 1.
 - (б) Изабрати реч из теме која је одабрана у 2.а). Како свака тема има речи које фаворизује (веће вероватноће у односу на остале речи), то ће се те речи највероватније одабрати у овом кораку.

Описана процедура може се графички илустровати претходном сликом (2.1). Одабрана расподела тема у документу (тачка 1 описане процедуре) представљена је хистограмом на десној страни слике. Обојени кругови представљају одабир теме из документа (корак 2.а)) док речи повезане стралицама са њима представљају одабирану реч из те теме (корак 2.б)).¹

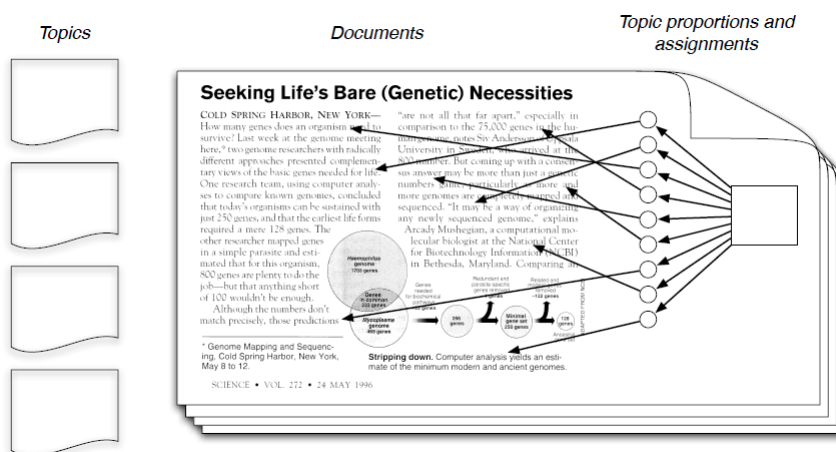
Формално, тема се дефинише као расподела речи над неким фиксним скупом речи - речником. Рецимо, тема биологија ће са већом вероватноћом садржати речи везане за ту област нпр. биљка, животиња, ћелија, ген итд. док ће тема везана за математику ове речи садржати са нижим вероватноћама у односу на речи нпр. број, разломак, променљива, коцка итд.

Дакле, основна карактеристика LDA-а је то што сви документи деле **исти скуп тема** али сваки документ те теме садржи у различитим односима. Овакаво посматрање докумената јако је природно и интуитивно.

2.1.2 Како се откривају теме - упрошћен пример

Проликом генерисања поменутог документа, било је познато која тема садржи које речи као и у којим односима је заступљена свака тема у тексту. Циљ алгоритама за моделовање тема је да **аутоматски** "открије" које су то теме присутне у неком документу и које речи припадају којој теми али **само** на основу речи које се јављају у документу, без било каквог додатног знања. Сазнање о томе која тема се у којој мери налази у неком документу, није од превеликог практичног значаја. Међутим, уколико је на располагању огромна количина докумената (нпр. дигитална база свих издања листа "Политика"), откривање сродних докумената, или докумената који се баве само одређеним темама може бити јако важно. Због тога ће се надаље говорити о скупу докумената над којим се извршава моделовање тема, уместо о једном документу. Притом, наравно, и даље важи претпоставка да сваки документ "говори о" свим темама које се могу издвојити из свих докумената, само у различитим односима.

Према свему реченом, једино што је *видљиво*, енг. **observed** су документи, односно речи које се у документима јављају. Тематска расподела по документима, као и расподела речи по темама су *скривене или невидљиве*, енг. *non observed, hidden* (Слика 2.2)



Слика 2.2: Пример чланка, преузето са [3]

Основни задатак алгорита је окривање скривених структура на основу видљивих. У овом тренутку, моделовање тема можемо посматрати као **обрнути** генеративни процес.

¹Расподела на основу које се одабирају пропорције тема у кораку 1, назива се Дирихлеова расподела, енг. Dirichlet distribution. На основу те одабране расподеле, врши се придруживање речи документима, енг. allocate

Дакле, циљ моделовања тема је откривање скривених структура из којих су **највероватније**, генеративним процесом, добијени видљиве структуре, тј. документи. Током рада, откривају се удели различитих тема по документима као и расподеле речи унутар тема. Важно је напоменути да **именовање** тема не постоји у основној верзији алгоритама. Алгоритам групише речи у одређене целине - теме, а насловљавање тема се препушта стручњацима.

Нека је дат једноставан документ који, након склањања везника, личних заменица и осталих шумава (енг.*stopwords*) садржи речи приказане у следећој табели.

Etruscan	trade	price	temple	market

Слика 2.3: Преузето са [4]

Процес моделовања тема започео **случајним додељивањем** тема свакој од речи у документу. Дакле, пошто не постоји никакво знање о присуству тема у документима као ни о томе која реч припада којој теми, ову доделу је неопходно урадити на случајан начин. Интуитивно је јасно да се за тако нешто унапред мора одредити број тема који се захтева у задатом скупу докумената. Више о улазним параметрима алгоритама може се наћи у одељку 3 овог рада. Пример једне случајне доделе дат је на следећој слици .

3	2	1	3	1
Etruscan	trade	price	temple	market

Слика 2.4: Преузето са [4]

На овај начин је направљена иницијална **расподела** тема унутар посматраног документа - 40% текста говори о теми 3, 40% о теми 1, док 20% говори о теми 2.

Уколико сада на сличан начин доделимо теме и осталим документима, полазни скуп докумената може се приказати следећом сликом.

trade	ship	2	3	2	1	1	Italy	temple	ship	trade	market
-------	------	---	---	---	---	---	-------	--------	------	-------	--------

Слика 2.5: Преузето са [4]

Пошто сваки документ има иницијалну, **случајну** расподелу тема, једноставно је груписати речи унутар тема и на тај начин направити иницијалну **случајну** расподелу речи по темама. Одређивање расподеле по темама, може се илустровати следећом сликом.

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	8	1
...			

Слика 2.6: Преузето са [4]

У првој колони табеле уписане су све речи из свих докумената (речник) док је затим у свакој од наредних колона уписан одговарајући број који означава колико пута је дата реч додељена тој теми. Рецимо, бројеви 1 0 35 у другој врсти табеле означавају да је реч *Etruscan* 35 пута била сврстана у тему број 3, једном је била додељена теми број 1 док се ниједним није нашла у теми број 2.. Дакле, почевши од друге колоне приказане табеле, табела по колонама, садржи *расподелу речи по темама*. Једноставним сортирањем колона, добијају се највероватније речи у свакој од тема.


У овом тренутку, расподеле које су добијене нису релевантне зато што у позадини стоји апсолутно случајно додељивање тема које није базирано на документима, тј. на јединим видљивим подацима. Дакле, потребно је добијене резултате *прилагодити* тако да осликавају тематску структуру документа. Прилагођавање се одвија у одређеном, унапред познатом броју итерација. Генерално, што је већи број итерација, то су добијене расподеле релевантније, мада, како резултати показују, постоје и нека ограничења за ове вредности. Више о одређивању оптималног броја итерације може се наћи у одељцима 3 и 6 овог рада.

Унутар једне итерације, за сваки документ за сваку реч унутар тог документа врши се провера колико је тренутно додељена тема адекватна, тј. да ли постоји боља тема којој би та реч могла бити додељена. На тај начин, из итерације у итерацију, расподеле све више и више осликавају структуру докумената.

Нека је почетна расподела по темама дата на претходним сликама (2.5,2.6). Нека се провера подобности теме прво извршава за реч *trade* првог документа. Ова реч је унутар првог документа додељена теми број 2, док је, гледано са становишта свих докумената који се посматрају, укупно 8 пута сврстана у ту тему. Потребно је испитати да ли тема број 2 највише одговара тој речи. Претпоставимо да знамо теме за све остале речи, како из документа који се посматра, тако и за остале документе, и да је једино непознато којој теми припада *trade* у посматраном документу. Дакле, расподела тема у посматраном документу као и расподела речи по темама, сада изгледа као на следећој слици и потребно је доделити изабраној речи тему унутар посматраног документа.

3	?	1	3	1
Etruscan	trade	price	temple	market

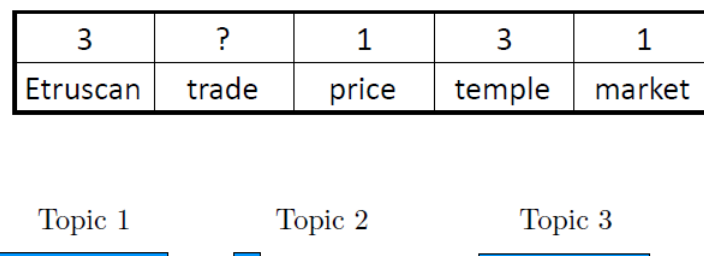
	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	10	7	1
...			



Слика 2.7: Преузето са [4]

Уколико се посматра расподела тема у изабраном документу, приметиће се да документ највише "говори о" темама 3 и 1 док о теми 2 не говори уопште. Према томе, удели тема 3 и 1 су значајни, док је удео теме 2 занемарљиво мали. Обзиром да је основна претпоставка овог модела да сви документи говоре о свим темама, не може се рећи да изабрани документ уопште "не говори" о теми 2. Начин на који ће се означити да је тема 2 јако слабо присутна у изабраном документу врши се тако што се теми 2 додели јако мали удео. Обзиром да је у питању реч из документа који има неки одређену расподелу тема, логично је очекивати да избор теме за ту реч зависи од тема тог документа.

Удели тема у изабраном документу могу се представити дужином линија, као што је приказано на следећој слици.



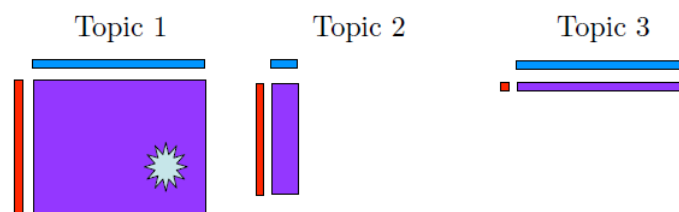
Слика 2.8: Преузето са [4]

Међутим, како је речник (скуп речи) заједнички за све документе, тема речи зависи и од глобалног присуства те речи у свим темама. И ова претпоставка је логична, јер, као што је на почетку наведено, свака реч припада свим темама, само са различитим вероватноћама. Глобално присуство изабране речи у свим темама приказано је на следећој слици :

	1	2	3
trade	10	7	1

Слика 2.9: Преузето са [4]

Дакле, избор теме за реч *trade* зависи од расподеле тема у посматраном документу као и од присуства те речи у свим темама. Ова зависност може се представити следећом сликом :



Слика 2.10: Преузето са [4]

Вертикална, црвена линија представља присуство речи у одговарајућој теми (формалније, вероватноћу са којом се та реч налази у изабраној теми). Љубичаста "површина" представља подобност да одговарајућа тема буде додељена тој речи у изабраном документу. Како се слике јасно може уочити, "највећу" површину формира тема 1 те је, према томе, речи *trade* додељује тема број 1 у овој итерацији. Након извршене измене, расподела тема у документу, као и расподела речи по темама, приказана је на следећој слици :

3	1	1	3	1
Etruscan	trade	price	temple	market

	1	2	3
Etruscan	1	0	35
market	50	0	1
price	42	1	0
temple	0	0	20
trade	11	7	1
...			

Слика 2.11: Преузето са [4]

На овај начин, изабраној речи додељена је највероватнија тема а тематска слика документа "више личи" на реалну слику.

Ако се описани процес примени на сваку реч сваког документа, расподеле ће из итерације у итерацију све више осликавати структуру полазних докумената. Обзиром да избор теме за сваку реч зависи од тренутно додљених расподела тема и речи, овим процесом се узимају у обзир видљиви подаци. Дакле, скривене структуре (непознате расподеле) се генеришу на основу једино видљивих података - докумената и речи у њима.

Узевши све претходно у обзир, јасно је због чега не постоји именовање тема. Када се процес "генерисања" заврши (нпр. достигне одређени број итерација), сортирањем одговарајућих колона из табеле са слике 2.7, добија се расподела речи по темама. На основу експертског знања овим расподелама додељују се имена - нпр. тема 1 - математика, тема 2 - економија итд.

2.1.3 Графички пример моделовања тема

Алгоритми за моделовање тема, поред текста, могу се применити и на друге облике података. Једна од могућих примена је и "октирвање тематике" слика. Рецимо, помоћу ове групе алгоритама могуће је препознати "сличне" слике. Овај рад неће се бавити овом врстом примене, али се ТМ алгоритама може јако добро илустровати једноставним графичким примером. Ради бољег разумевања шта ТМ уствари ради, биће описан један пример примене ТМ над сликама.

Глава 3

Математичка позадина

У претходним поглављима, рад се углавном бавио питањима *шта је ТМ алгоритам и чему служи*, без улажења у то **како** он уствари ради.

Опис рада ТМ алгоритама - конкретно LDA имплементације, биће изложен у неколико целина. Најпре ће се објаснити (увести) неки појмови вероватноће који су битни за разумевање суштине рада алгорита, а затим ће бити изнешена математичка позадина самог алгорита.

3.1 Теорија вероватноће

Теорија вероватноће је математичка дисциплина која се бави изучавањем случајних појава тј. појава чији исходи нису увек строго дефинисани.

Први проблеми који се могу сматрати проблемима вероватноће потичу још из 12. века и везани су за проучавање исхода разних игара на срећу. Развој *теорије вероватноће* почиње средином 17. века и везан је за имена Блеза Паскала, Пјера де Ферма и Кристијана Хајгенса. Наиме, између Паскала и Ферма је 1654. године започела интересантна преписка о низи проблема међу којима је био и проблем везан за поделу улога приликом прекида једне коцкарске игре. Проблем је био постављен на следећи начин : Два играча А и Б се договоре да читав улог припадне ономе ко први добије три игре. Када је играч А добио 2 игре а играч Б једну игру, играчи су споразумно одлучили да прекину игру. Поставља се питање како сада да поделе улог. Паскал је предложио поделу у односу 3:1 у корист играча А. Овај пример често се узима као почетак настанка теорије вероватноће.

Неке од појава које се догађају у реалном свету лако се могу предвидети и објаснити услед познавања законитости њиховог настанка. У такве појаве спадају нпр. помарачење Сунца и Месеца, плима и осека, гравитација итд. Међутим, постоје појаве чије узроке тренутно није могуће одредити па се не могу у потпуности објаснити и одредити. Неке од таквих појава су нпр. добитак на лутрији или метеоролошке појаве. Прилоком бацања металног, хомогеног новчића, никада није сигрно да ли ће пасти писмо или глава. Међутим, уколико бацамамо новчић много пута, може се уочити да је отприлике исти број пута пало писмо као и глава (такве експерименте су радили Буфон и Пирсон Дакле, законитост код оваквих догађаја може се уочити тек након великог броја понављања појаве.

3.1.1 Основни појмови

Основни полазни појам у теорији вероватноће је непразан скуп Ω који представља скуп свих могућих исхода једног експеримента. Овај скуп се често назива **простор елементарних догађаја** и може бити коначан, пребројив или непребројив. **Случајни догађаја** или само **догађај** представља било који подскуп скупа Ω . Најчешће се случајни догађаји означавају великим, штампаним, латиничним словима. За догађај **A** каже се да се **реализовао** ако

се реализовао неки исход ω који припада скупу A . Догађај који је садржи све могуће елементарне исходе експеримента назива се **сигуран догађај** а догађај који не садржи ни један елементарни исход назива се **немогућ догађај**.

Пример: Нека је дата хомогена коцка чије су стране означене бројевима од 1 до 6. Елементарни догађаји су појављивање одређеног броја при бацању коцкице. Према томе, скуп свих могућих исхода експеримента бацања коцкице је $\Omega = \{1, 2, 3, 4, 5, 6\}$. Догађај $A =$ "пао је паран број" одређује скуп $A = \{2, 4, 6\}$

Производ два догађаја и, у ознаци AB је догађај који се реализује ако и само се ако реализују оба догађаја. Дакле, производ догађаја је пресек скупова A и B . Уколико су A и B дисјунктни скупови (пресек је празан скуп) за такве догађаје кажемо да су **несагласни** или да се **искључују**.

Збир два догађаја A и B , у ознаци $A \cup B$ представља догађај који се реализује ако се реализује бар један од догађаја A и B .

Разликом догађаја A и B , у ознаци $A - B$ назива се догађај који се реализује ако и само ако се реализује догађај A а не реализује догађај B .

Потпун систем догађаја : За догађаје A_1, A_2, \dots, A_n се каже да образују *потпун систем догађаја* уколико важи : $\bigcup_i A_i = \Omega$. Дакле, при реализацији неког експеримента бар један од догађаја A_1, A_2, \dots, A_n ће се реализовати. Посебно су интересантни потпуни системи несагласних догађаја као што се може видети код формуле тоталне вероватноће.

Дефиниција 1 (Класична дефиниција вероватноће) : Нека је $\Omega = \{\omega_1, \omega_2, \dots, \omega_n\}$ скуп свих могућих једнаковероватних елементарних догађаја који су међусобно несагласни и нека је $A = \{\omega_{i_1}, \omega_{i_2}, \dots, \omega_{i_m}\}$ догађај који се састоји од m елементарних једнаковероватних догађаја. Вероватноћа наступања догађаја A је :

$$P(A) = \frac{m}{n} \quad (3.1)$$

Претходна дефиниција може се неформално изразити и овако : вероватноћа догађаја A једнака је количнику броја **повољних исхода** експеримента (исходи када се реализује догађај A) и укупног броја свих могућих исхода експеримента.

Класична дефиниција вероватноће је применљива само онде где су елементарни догађаји једнаковероватни. Међутим, тај услов је у пракси јако тешко испунити. Чак и у случајевима када је то наизглед очигледно, као што је бацање коцкице, једнаковероватност не може бити гарантована. Разлози за то могу бити технологија израде коцкице која не мора бити савршено прецизна, немогућност обезбеђивања идеалних и непромењљивих услова током извођења експеримента итд. Због тога је једини начин којим је могуће заиста утврдити вероватноћу догађаја A *статистички приступ* заснован на великом броју експеримената.

Дефиниција 2 (Статистичка дефиниција вероватноће) : Нека се у n понављања експеримента изведених под приближно истим условима догађај A реализовао m_n пута. Вероватноћа догађаја A је

$$P(A) = \lim_{n \rightarrow \infty} \frac{m_n}{n} \quad (3.2)$$

3.1.2 Условна вероватноћа

Вероватноћа догађаја чија реализација **не зависи** од наступања било ког другог догађаја назива се **безусловна вероватноћа**. Ако је реализација догађаја A условљена реализацијом неког догађаја B при чему B није немогућ догађај ($P(B) \neq 0$), тада се вероватноћа догађаја под условом да се десио догађај B назива **условном вероватноћом** и означава се са $P(A | B)$. Дакле, $P(A | B)$ је вероватноћа догађаја A под условима који сигурно доводе до реализације догађаја B .

Нека се изводи експеримент у коме постоји n једнаковероватних елементарних догађаја и нека је са n_A, n_B, n_{AB} означен број елементарних догађаја који доводе до реализације догађаја A, B, AB редом.

Према класичној дефиницији вероватноће, вероватноћа реализације догађаја A и AB је :

$$P(B) = \frac{n_B}{n}, P(AB) = \frac{n_{AB}}{n} \quad (3.3)$$

Ако је реализација догађаја A условљена реализацијом догађаја B , то је број повољних исхода догађаја A n_{AB} (број елементарних догађаја који имају осбине и скупа A и скупа B). Пошто се догађај A реализује само ако се реализовао догађај B , број свих могућих исхода је n_B (број свих могућих елементарних догађаја када наступа догађај B). Дакле, условна вероватноћа догађаја A , под условом да се десио догађај B је :

$$P(A | B) = \frac{n_{AB}}{n_B} = \frac{\frac{n_{AB}}{n}}{\frac{n_B}{n}} = \frac{P(AB)}{P(B)}, P(B) \neq 0 \quad (3.4)$$

У случају да је догађај B условљен догађајом A , аналогно се изводи да је

$$P(B | A) = \frac{P(AB)}{P(A)}, P(A) \neq 0 \quad (3.5)$$

Из релација (3.4) и (3.5) следи

$$P(AB) = P(B) \cdot P(A | B) = P(A) \cdot P(B | A) \quad (3.6)$$

Релација (3.6) назива се још и **теорема о производу вероватноћа**

Теорема 1 (Формула тоталне вероватноће) : Ако су H_1, H_2, \dots, H_n међусобно несагласни догађаји, $P(H_i) > 0 (i = 1, \dots, n)$ при чему важи $H_1 + H_2 + \dots + H_n = \Omega$ тада је :

$$P(A) = \sum_{i=1}^n P(H_i)P(A | H_i) \text{ за сваки догађај } A \subseteq \Omega \quad (3.7)$$

Напомена : Догађаји H_1, H_2, \dots, H_n чине потпун систем несагласних догађаја.

Доказ 1 Обзором да су догађаји подскупови скупа свих елементарних догађаја очигледно је да важи

$$A = A\Omega = A \sum_{i=1}^n H_i = \sum_{i=1}^n AH_i. \quad (3.8)$$

На основу релације (3.6) следи :

$$P(A) = P\left(\sum_{i=1}^n AH_i\right) = \sum_{i=1}^n P(AH_i) = \sum_{i=1}^n P(H_i)P(A | H_i) \quad (3.9)$$

Вероватноће $P(H_i)$ су обично познате унапред и називају се **априорним вероватноћама** а сами догађаји **хипотезама**.

Теорема 2 (Бајесова формула ¹) : Ако су H_1, H_2, \dots, H_n међусобно несагласни догађаји, $P(H_i) > 0 (i = 1, \dots, n)$ при чему важи $H_1 + H_2 + \dots + H_n = \Omega$ тада је :

$$P(H_i | A) = \frac{P(H_i)P(A | H_i)}{\sum_{j=1}^n P(H_j)P(A | H_j)} \quad (i = 1 \dots n) \quad \text{за сваки догађај } A \subseteq \Omega \quad (3.10)$$

Доказ 2 Из релације (3.6) следи :

$$P(H_i A) = P(H_i)P(A | H_i) = P(A)P(H_i | A) \quad (i = 1...n) \quad (3.11)$$

Условна вероватноћа догађаја H_i под условом да се десио догађај A је:

$$P(H_i | A) = \frac{P(H_i A)}{P(A)} = \frac{P(H_i)P(A | H_i)}{P(A)}$$

Примењујући формулу потпуне вероватноће за $P(A)$ добија се

$$P(H_i | A) = \frac{P(H_i)P(A | H_i)}{\sum_{j=1}^n P(H_j)P(A | H_j)}$$

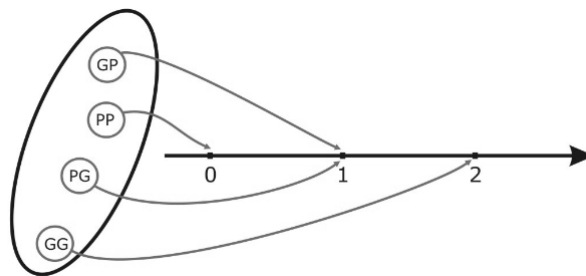
што представља Бајесову формулу.

3.1.3 Случајне променљиве

Ако се сваком елементарном догађају придружи један реалан број, онда се извођење експеримента може посматрати као избор вредности једне променљиве. Променљива величина која те бројене вредности узима са одређеним вероватноћама назива се *случајна променљива*. Дакле, уместо вербалне карактеризације догађаја (описа речима шта догађај представља) много је једноставније за рад догађаје окарактерисати бројним вредностима тј. неким реалним бројевима.

Пример 1 : У експерименту бацања новчића могућа су два елементарна исхода : грб или писмо. Нека је догађај који се посматра "пало је писмо". Појава писма се може означити бројем 1 а појава грба бројем 0. Сада се овај експеримент може замислити као избор 0 или 1 са вероватноћом $\frac{1}{2}$

Пример 2 : Новчић се баца два пута. Нека је са P означена појава писма а са G појава грба. Скуп свих елементарних исхода експеримента је $\Omega = \{PP, PG, GP, GG\}$. Нека је догађај који се посматра "број палих писама". Сваком исходу се може доделити један реалан број и то $PP \rightarrow 2, GP \rightarrow 1, PG \rightarrow 1, GG \rightarrow 0$. Ово додељивање вредности се карактерише случајном променљивом. Случајна променљива сваку од ових вредности узима са различитом вероватноћом.



Слика 3.1: Графички пример случајне променљиве

Дефиниција 3 Функција X која сваком случајном догађају $\omega \in \Omega$ додељује неки реалан број $X(\omega)$ назива се *случајна променљива* где је $X : \Omega \rightarrow R$

Дакле, случајна променљива је пресликавање скупа Ω у скуп **реалних** бројева за разлику од вероватноће која је пресликавање скупа Ω у скуп $[0, 1]$

Важно је уочити да случајна променљива **нема конкретну вредност** већ се само говори о вероватноћама да узме неки конкретну вредност.

Разликују се два основна типа случајних променљивих - **дискретне** и **непрекидне**. Подела се врши у зависности од тога да ли случајна променљива узима вредности у пребројивом или непребројивом скупу вредности.

3.1.3.1 Дискретне случајне променљиве

За случајну променљиву се каже да је дискретног типа ако узима коначан број изолованих вредности или пребројиво много вредности

Дефиниција 4 Нека случајна променљива X може да узме вредности x_1, x_2, \dots, x_n са вероватноћама p_1, p_2, \dots, p_n при чему важи да је $p_1 + p_2 + \dots + p_n = 1$. Скуп парова $(x_i, p_i = P\{X = x_i\})$, $i = 1, 2, \dots, n$ или написано :

$$\begin{pmatrix} x_1 & x_2 & \cdots \\ p(x_1) & p(x_2) & \cdots \end{pmatrix}$$

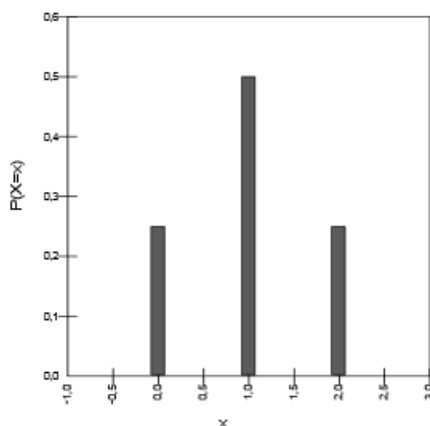
чине закон расподеле или распоред вероватноћа случајне променљиве X .

Закон расподеле случајне променљиве може да се посматра као **правило** по коме се свакој вредности случајне променљиве придружује одговарајућа вероватноћа. Дакле, при реализацији експеримента сигурно ће се десити догађај којем је придружена нека вредност случајне променљиве. Због тога је сума свих вероватноћа у расподели случајне променљиве 1. Међутим, нису све вредности подједнако вероватне па се свакој вредности придружује вероватноћа са којом се очекује. Претходна дефиниција може се интерпретирати и на следећи начин : извесна маса једнака јединици је распоређена на такав начин да се у тачкама x_1, x_2, \dots, x_n налазе одговарајући делови масе p_1, p_2, \dots, p_n . Услед оваквог тумачења, закон расподеле вероватноћа се често назива и **функција масе вероватноћа**

У примеру 2, случајна променљива може да узме три вредности, тј. писмо се може појавити 0, 1 или 2 пута у два бацања. Ни један други исход није могућ - нпр. у два бацања писмо не може да се појави 3 пута или -1 пут. Међутим, вероватноћа да се писмо неће појавити ни једном (или да се појави два пута) је $\frac{1}{4}$ - вероватноћа да падне глава у првом бацању је $\frac{1}{2}$ и вероватноћа да падне глава у другом бацању је $\frac{1}{2}$, дакле, вероватноћа да оба пута падне глава је $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$, вероватноћа да се писмо појави тачно једном је $\frac{1}{2}$ - писмо пада тачно једном у случају PG или GP. Вероватноћа за оба ова догађаја је $\frac{1}{4}$. Дакле, вероватноћа да се десио бар један од ових догађаја је $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$. Према томе, расподела случајне променљиве "број појављивања писма у два бацања" је :

$$\begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix}$$

Закон расподеле дискретне случајне променљиве може се представити графички, као на следећој слици :



Слика 3.2: Графички пример случајне променљиве

На апсциси се налазе могуће вредности случајних променљивих док се на ординати налазе вероватноће са којом случајна променљива узима дату вредности. Са претходне слике јасно се уочава дискретност посматране случајне променљиве - вероватноћа да случајна променљива узме вредност између неке две целобројне вредности је 0.

Функција расподеле дискретне случајне променљиве:

Распоред или закон расподеле случајне променљиве дискретног типа може се представити као листа свих могућих вредности случајне променљиве и одговарајућих вероватноћа. Међутим, поставља се питање како представити случајну променљиву која може узимати јако пуно вредности тј. бесконачно много вредности. У овом случају би требало формирати листу од бесконачно много чланова, што је практично неизводљиво. (Пример једне такве случајне променљиве би био - број бацања коцкице док се не добију две узастопне шестике. Случајна променљива може узети вредности 2,3,4,... са различитим вероватноћама, при чему не постоји горња граница броја бацања при којој се сигурно добијају две узастопне шестике). Због описаног проблема, потребно је пронаћи другачији начин представљања случајне променљиве и одговарајућих вероватноћа. То се постиже **функцијом расподеле** која се може дефинисати за сваку случајну променљиву.

Дефиниција 5 Функција расподеле (још се назива и **кумулативна функција расподеле**) дискретне случајне променљиве претставља вероватноћу да случајна променљива X узме вредност која је мања или једнака неком реалном броју x при чему је дефинисана за свако реално x .

$$F(x) = P(X \leq x) \quad \forall x \in R$$

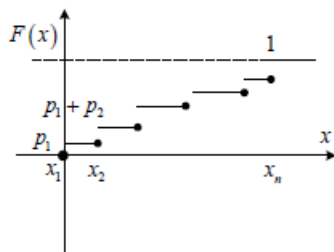
Дакле, кумулативна функција расподеле има облик

$$F(x) = \begin{cases} 0, & x \leq x_1 \\ p_1, & x_1 < x \leq x_2 \\ p_1 + p_2, & x_2 < x \leq x_3 \\ \dots & \dots \\ 1 & x > x_n \end{cases}$$

и може се изразити као :

$$F(x) = \sum_{k, x_k \leq x} P(X = x_k)$$

а графички приказ је дат на следећој слици :



Слика 3.3: График кумулативне функције расподеле случајне променљиве дискретног типа

Две најважније дискретне расподеле су **Биномна** и **Пуасонова** расподела.

3.1.3.2 Непрекидне случајне променљиве

Случајна променљива је (апсолутно) непрекидног типа ако може да узме **било коју** вредност из неког интервала. Број вредности које може да узме случајна променљива непрекидног типа је **бесконачан**. Неки од примера су : висина и тежина људи, дужина трајања батерије итд. На пример, нека је X случајна променљива која представља дужину рада сијалице. Ова случајна променљива може узети било коју вредност на интервалу од 1 до нпр. 1000 сати. Како у интервалу $[0, 1000]$ има бесконачно много реалних бројева, не постоји начин да се дефинише вероватноћа за сваку појединачну вредност, као што је био случај код дискретних променљивих. Такође, интуитивно је јасно да је вероватноћа да ће сијалица прегорети у тачно одређеном тренутку једнака 0 док је вероватноћа да ће прегорети у неком временском интервалу различита од нуле.

Дефиниција 6 Случајна променљива X је апсолутно непрекидног типа ако постоји **ненегативна** функција $f: \mathbb{R} \rightarrow \mathbb{R}$ таква да за било који интервал $[a, b] \subset (-\infty, \infty)$ важи :

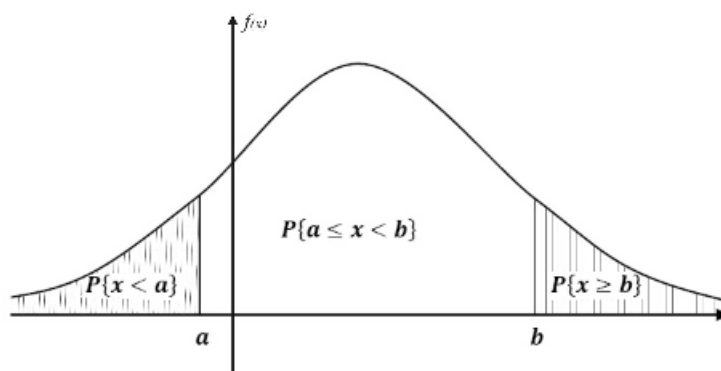
$$P\{a \leq X < b\} = \int_a^b f(x)dx \quad (3.12)$$

Функција $f(x)$ мора да задовољи услов :

$$P\{-\infty \leq X < \infty\} = P\{\Omega\} = \int_{-\infty}^{\infty} f(x)dx = 1$$

Функција $f(x)$ се назива **густина расподеле вероватноће** случајне променљиве X . Случајне променљиве **дискретног типа** немају густину расподеле баш као што ни случајне променљиве непрекидног типа немају закон расподеле вероватноћа.

Из релације (3.12) следи да је вероватноћа да случајна променљива узме вредност из интервала $[a, b]$ једнака **површини** испод графика функције $f(x)$ на интервалу $[a, b]$.



Слика 3.4: Функција густине

Функција расподеле непрекидне случајне променљиве:

Дефиниција 7 *Функција расподеле (кумулативна функција расподеле) непрекидне случајне променљиве се може представити као :*

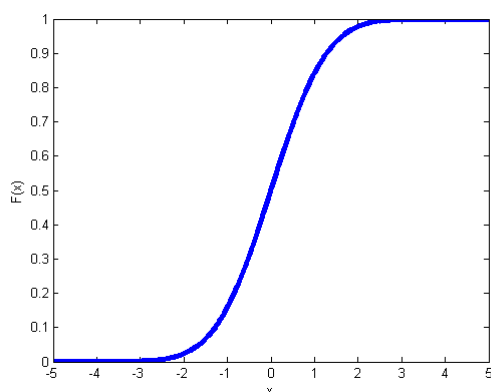
$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt \quad x \in (-\infty, \infty)$$

где је $f(x)$ функција густине.

Дефиниција кумулативне суме преко интеграла је јаснија ако се има на уму интервал из ког случајна променљива може да узме вредности. Код случајних променљивих дискретног типа, тај скуп је био пребројив па се кумулативна функција расподеле дефинисала преко суме. Случајне променљиве непрекидног типа могу узети бесконачно много вредности па се сума код дискретних случајних променљивих (када број тачака тежи у бесконачност) замењује интегралом.

Напомена : Ако случајна променљива X не узима све вредности из интервала $(-\infty, \infty)$ усваја се да је $f(x) = 0$ за све вредности x из интервала у којима X не узима вредности.

График кумулативне функције расподеле непрекидне случајне променљиве X је сада представљен глатком кривом линијом (за разлику од случајне променљиве дискретног типа где је график био "степенаст").



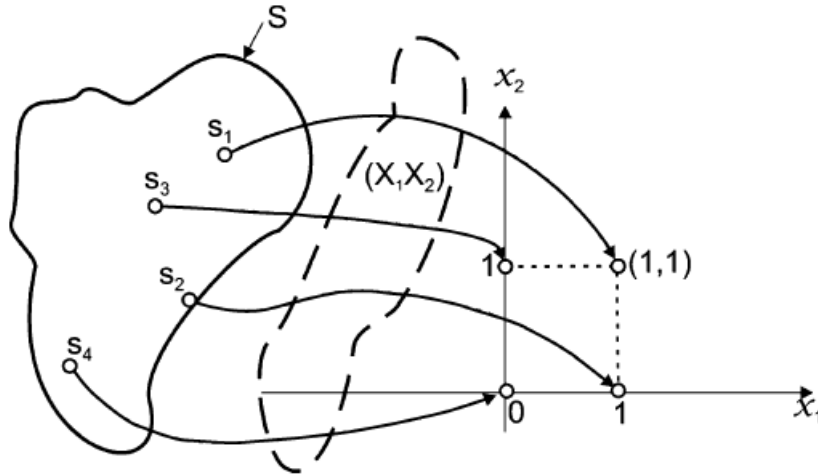
Слика 3.5: Кумулативна функција расподеле за случајне променљиве непрекидног типа

3.1.3.3 Вишедимензионалне случајне променљиве

Случајна променљива представља пресликавање скупа догађаја у реалне бројеве. Дакле, излази експеримента се мапирају у једнодимензионалан простор реалних бројева. Међутим, постоје случајеви када је потребно излазе експеримента мапирати у вишедимензионалне реалне просторе. На пример, при истовременом бацању два новчића могућа су 4 исхода :

1. s_1 : први новчић писмо - други новчић писмо
2. s_2 : први новчић писмо - други новчић глава
3. s_3 : први новчић глава - други новчић писмо
4. s_4 : први новчић глава - други новчић глава

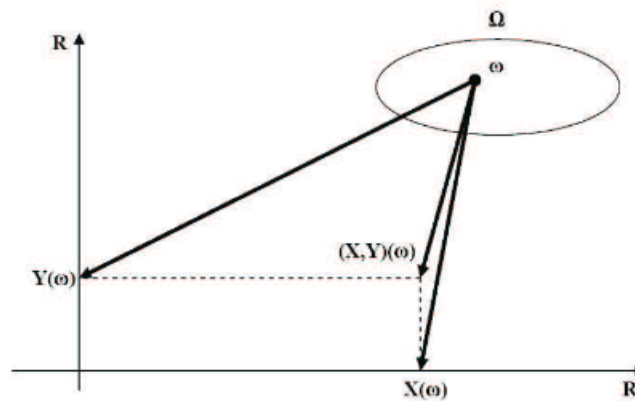
Нека је са X_1 означена случајна променљива која узима вредност 1 ако се на првом новчићу појавила глава, односно 0 ако се појавило писмо и аналогно X_2 која на исти начин означава појаву главе на другом новчићу. Исход експеримента се сада може описати дводимензионалном променљивом (X_1, X_2) . Графички приказ ове дводимензионалне променљиве дат је на следећој слици :



Слика 3.6: Експеримент : бацање два новчића.
Скуп S представља скуп свих елементарних исхода (Ω)

Дефиниција 8 Ако су $X : \Omega \rightarrow \mathbb{R}, Y : \Omega \rightarrow \mathbb{R}$ случајне променљиве, тада се **уређени пар** (X, Y) назива **двостепенациона случајна променљива**. Уређеним паром (X, Y) се сваком исходу $\omega \in \Omega$ придружује уређени пар бројева $(X(\omega), Y(\omega)) = (x, y) \in \mathbb{R} \times \mathbb{R} = \mathbb{R}^2$.

На следећој слици графички је представљена двостепенациона случајна променљива.



Слика 3.7: Двостепенациона случајна променљива

Овако уведен појам двостепенационале случајне променљиве се може проширити и на више димензија и тада настају n -димензионалне случајне променљиве. Закључци изведени за двостепенационалне се такође односе и на вишестепенационалне случајне променљиве.

Кумулативна функција расподеле (још се назива и *заједничка расподела* енгл. *joint distribution* двостепенационале случајне променљиве, у ознаци $F_{XY} : \mathbb{R}^2 \rightarrow [0, 1]$ дефинише се као вероватноћа реализације догађаја $\{X \leq x, Y \leq y\}$ односно :

$$F_{X,Y}(x, y) = P\{X \leq x, Y \leq y\} \quad -\infty < x, y < \infty$$

Неке карактеристике функције расподеле двостепенационале случајне променљиве :

1. $0 \leq F_{X_1, X_2}(x_1, x_2) \leq 1$
2. $F_{X_1, X_2}(-\infty, -\infty) = 0$

$$3. F_{X_1, X_2}(-\infty, -\infty) = 0$$

$$F_{X_1, X_2}(-\infty, x_2) = 0 \quad F_{X_1, X_2}(x_1, -\infty) = 0$$

$$4. F_{X_1, X_2}(\infty, \infty) = 1$$

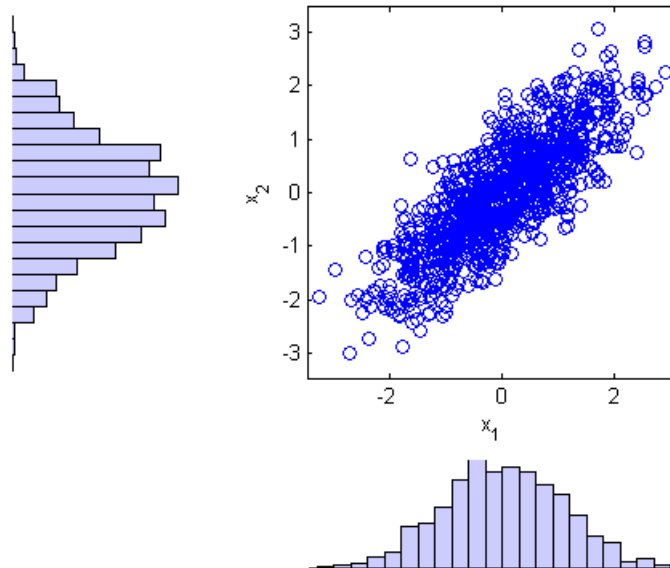
5.

$$F_{X_1, X_2}(x_1, \infty) = F_{X_1}(x_1) \quad (3.13)$$

6.

$$F_{X_1, X_2}(\infty, x_2) = F_{X_2}(x_2) \quad (3.14)$$

Једнакостима (3.13) и (3.14) су дефинисане **маргиналне расподеле** случајних променљивих X_1 и X_2 . Маргиналне расподеле су уствари расподеле једнодиментионалних случајних променљивих X_1 и X_2 . На следећој слици је представљена заједничка расподела две случајне променљиве заједно са њиховим маргиналним расподелама.

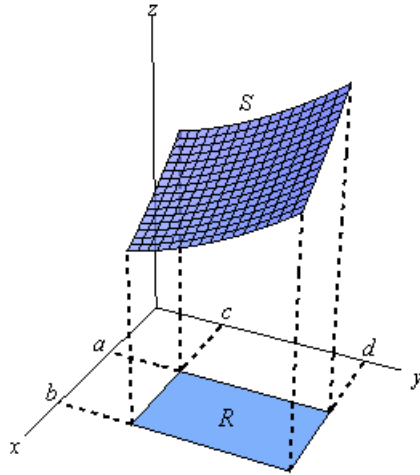


Слика 3.8: Двостепенационална случајна променљива и маргиналне расподеле

Заједничка функција густине (енг. joint density function) двостепенационална случајне променљиве се дефинише као :

$$f_{X_1, X_2}(x_1, x_2) = \frac{d^2 F_{X_1, X_2}(x_1, x_2)}{dx_1 dx_2} \quad (3.15)$$

У случају једнодиментионалне случајне променљиве, површина испод графика функције густине на неком интервалу представља је вероватноћу да случајна променљива узме вредност из тог интервала. У случају двостепенационалне случајне променљиве од интереса је пронаћи вероватноћу да она узме вредност из неке **области**. Та вероватноћа представља **запремину** тела ограниченог функцијом густине са горње стране и датом облашћу са доње стране.



Слика 3.9: Вероватноћа да дводемпзионална случајна променљива (X, Y) узме вредности из области R

Неке особине функције густине :

$$1. \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X_1, X_2}(x_1, x_2) dx_1 dx_2 = 1$$

$$2. F_{X_1, X_2}(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_{X_1, X_2}(x_1, x_2) dx_1 dx_2$$

$$3. F_{X_1}(x_1) = \int_{-\infty}^{x_1} \int_{-\infty}^{\infty} f_{X_1, X_2}(x_1, x_2) dx_1 dx_2$$

$$F_{X_2}(x_2) = \int_{-\infty}^{x_2} \int_{-\infty}^{\infty} f_{X_1, X_2}(x_1, x_2) dx_1 dx_2$$

$$4. f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1, X_2}(x_1, x_2) dx_2$$

$$f_{X_2}(x_2) = \int_{-\infty}^{\infty} f_{X_1, X_2}(x_1, x_2) dx_1$$

$$5. P\{x_{11} < X_1 \leq x_{12}, x_{21} < X_2 \leq x_{22}\} = \int_{x_{21}}^{x_{22}} \int_{x_{11}}^{x_{12}} f_{X_1, X_2}(x_1, x_2) dx_1 dx_2$$

Функција условне расподеле и густине

У неким специфичним случајевима је потребно пронаћи расподелу једне случајне променљиве знајући вредност друге случајне променљиве. Таква расподела назива се **условном расподелом** и обележава се са $F_{X_1}(x_1 | X_2 = x_2)$. Аналогно, може се дефинисати и проблем налажења функције густине једне случајне променљиве знајући вредност друге случајне променљиве и таква функција густине се означава са $f_{X_1}(x_1 | X_2 = x_2)$. Према [8] условна расподела односно густина се рачуна по следећем обрасцу (детаљно извођење се може наћи у [8])

$$F_{X_1}(x_1 | X_2 = x_2) = \frac{\int_{-\infty}^{x_1} f_{X_1, X_2}(x_1, x_2) dx_1}{f_{X_2}(x_2)}$$

односно :

$$f_{X_1}(x_1 | X_2 = x_2) = \frac{f_{X_1, X_2}(x_1, x_2)}{f_{X_2}(x_2)}$$

3.2 Важније расподеле

3.2.1 Биномна и полиномна(енг. multivariate) расподела

Биномна расподела

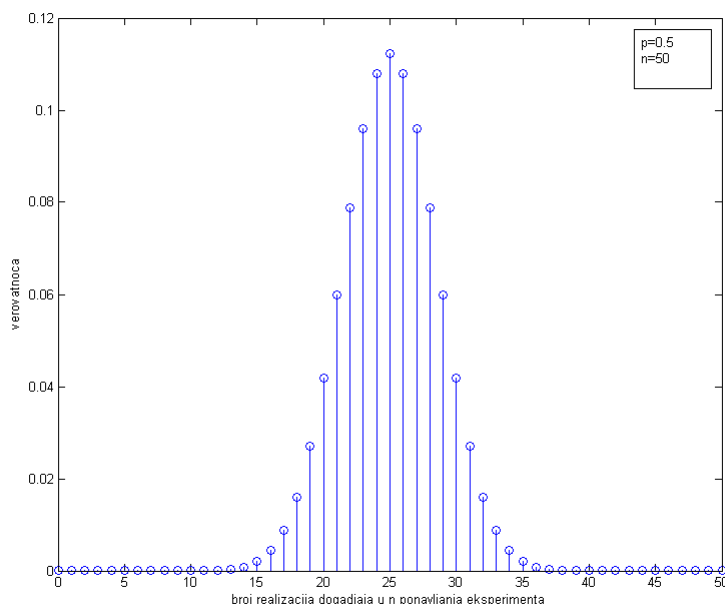
Нека ја A догађај неког експеримента E који се реализује са вероватноћом $P(A) = p$. Тада је вероватноћа супротног догађаја $P(\bar{A}) = 1 - p = q$. Резултат експеримента који је до интереса је остваривање или неостваривање догађаја A . Дакле, може се сматрати да је скуп свих елементарних исхода $\Omega = \{A, \bar{A}\}$. Нека се експеримент понавља **независно** и у неизмењеним условима n пута. На тај начин је формиран **сложени експеримент** чији скуп елементарних исхода садржи све могуће n -торке састављене од A и \bar{A} и има их укупно 2^n . Нека је, даље, на том скупу елементарних исхода дефинисана случајна променљива X_n као број остваривања догађаја A у n понављања експеримента E . Вероватноћа да ова случајна променљива узме конкретну вредност k је :

$$p_k = P\{X_n = k\} = \binom{n}{k} p^k q^{n-k}$$

Вероватноће $P\{X_n = k\}, (k = 0, 1, \dots, n)$ дефинишу **биномну расподелу**, у ознаци $\mathbb{B}(n, p)$. Ова расподела је дискретног типа а њена функција расподеле(кумулятивна) се може изразити као :

$$F(x) = \begin{cases} 0 & , x \leq 0 \\ \sum_{k=0}^r \binom{n}{k} p^k q^{n-k} & 0 < r < n \\ 1 & x > n \end{cases}$$

Закон расподеле вероватноћа случајне променљиве биномне расподеле приказан је на следећој слици :



Слика 3.10: Биномна расподела - закон расподеле

Полиномна (енг. multivariate) расподела

Изводи се серија од n независних експеримената при чему резултат експеримента може бити један од **коначно много** догађаја : $A_1, A_2, \dots, A_k, \sum_{i=1}^k A_i = \Omega, P(A_i) = p_i (i = 1, 2, \dots, k)$. Ако се дефинише k -диментионална случајна применљива $(S_n^{(1)}, \dots, S_n^{(k)})$, где $S_n^{(i)}$ представља број релаизација случајног догађаја A_i у n независних експеримената, тада важи :

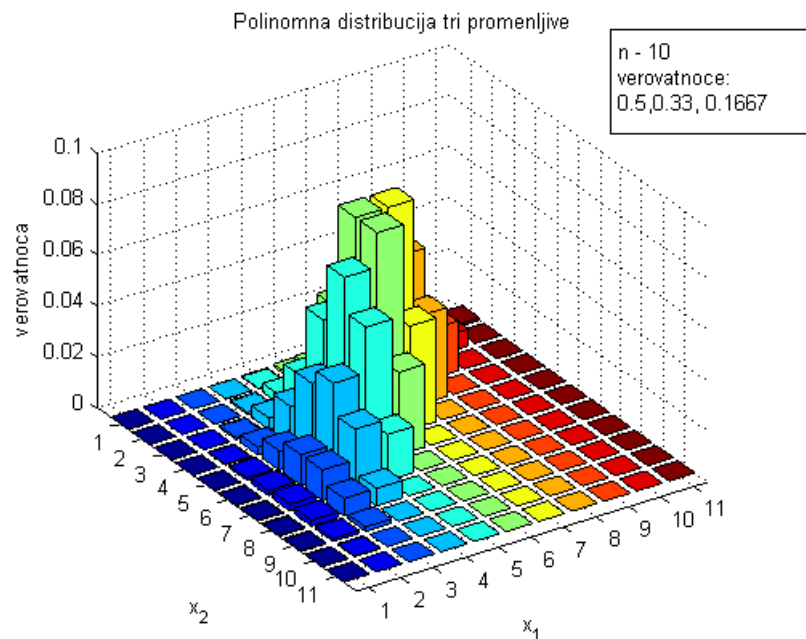
$$P(S_n^{(1)} = r_1, \dots, S_n^{(k)} = r_k) = \frac{n!}{r_1! \dots r_k!} p_1^{r_1} \dots p_k^{r_k}$$

$$r_1, \dots, r_k \in \{0, 1, \dots, n\} \quad r_1 + \dots + r_k = n$$

Ако се са $S = (S_n^{(1)}, \dots, S_n^{(k)})$ означи k -диментионална случајна применљива која има полиномијалну расподелу тада се то записује као :

$$S \sim Mult(n, p)$$

где је $p = (p_1, p_2, \dots, p_k)$ Пример полиномне расподеле при чему резултат експеримента може бити један од **три** догађаја, дат је на следећој слици :



Слика 3.11: Биномна расподела - закон расподеле

3.2.2 Дирихлеова расподела

Дирихлеова расподела представља фамилију расподела за параметре p полиномијалне расподеле. Задаје се са :

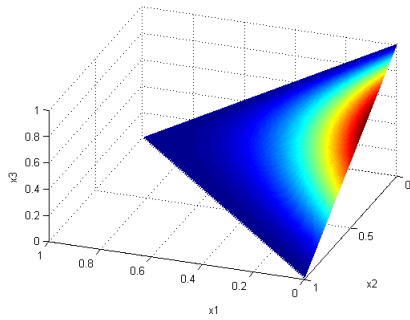
$$Dir(p; \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^K p_k^{\alpha_k - 1}$$

при чему је α параметар расподеле а B означава мултиномијалну бета функцију. Мултиномијалну бета функција се изражава преко гама функције на следећи начин

$$B(\alpha) = \frac{\prod_{i=1}^{|\alpha|} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{|\alpha|} \alpha_i)}$$

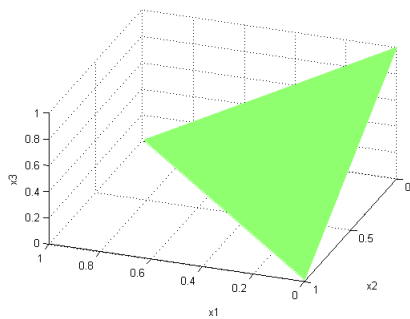
На слећој слици је графички представљена Дирихлеова расподела за три променљиве :

- $\alpha = (1, 2, 3)$



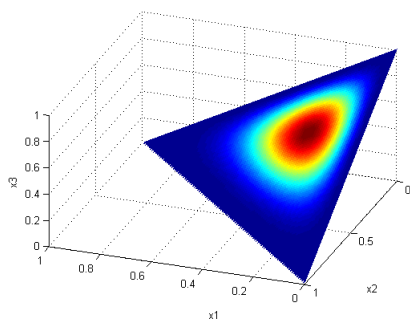
Слика 3.12: Дирихлеова расподела - интензивнија боја предтсваља већу вероватноћу

- $\alpha = (1, 1, 1)$

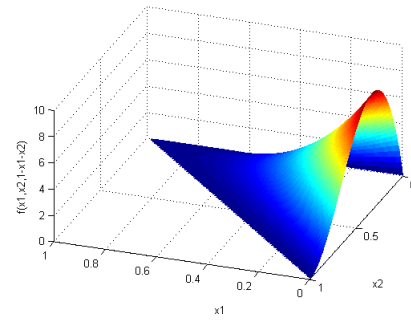


Слика 3.14: Дирихлеова расподела - интензивнија боја предтсваља већу вероватноћу

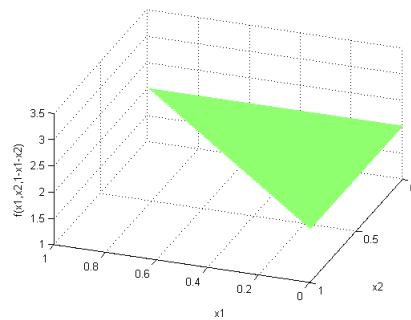
- $\alpha = (3, 3, 5)$



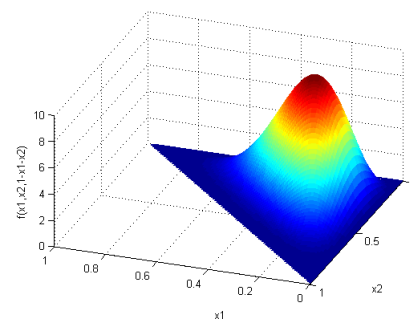
Слика 3.16: Дирихлеова расподела - интензивнија боја предтсваља већу вероватноћу



Слика 3.13: Дирихлеова расподела у три димензије



Слика 3.15: Дирихлеова расподела у три димензије



Слика 3.17: Дирихлеова расподела у три димензије

3.3 Гибсово узорковање

3.3.1 Марковљеви ланци

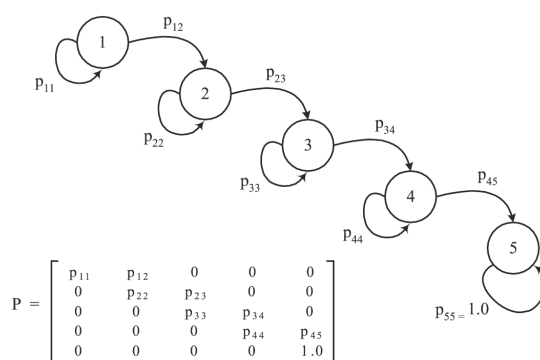
Марковљевим ланцима моделује се математички сиситем стања и прелаза међу тим стањима.

Дефиниција 9 *Случајан (стохастички) процес представља математички модел процеса чија је еволуција описана законима вероватноће.*

Марковљеви процеси су они случајни процеси чије будуће стање зависи само од тренутног стања. Оваква особина још се назива и одсуство памћења

Марковљеви ланци представљају посебну врсту Марковљевих процеса где се процес може налазити само у коначном броју стања.

Пример Марковљевог ланца дат је на следећој слици :



Слика 3.18: Марковљев ланац - графички пример и матрица транзиције

Систем се састоји од 5 стања. У сваком стању, са одређеним вероватноћама систем може да пређе у неко од следећих стања - конкретно да остане у тренутном стању или да пређе у једно стање ниже. Вероватноћа преласка у следеће стање зависи само од тренутног стања. Марковљеви ланци се често представљају **матрицама транзиције** при чему i -та врста у матрици садржи вероватноће преласка у свако од стање система када се систем налази у стању i . Сума свих вероватноћа у свакој врсти је 1 (систем сигурно мора да се нађе у неком стању, дакле вероватноћа да систем пређе у неко стање, могуће исто, је 1). Свака врста представља условни закон расподеле вероватноћа да систем пређе у било које стање у односу на тренутно стање (i -та врста - i -то стање). Свака колона представља маргиналну расподелу вероватноћа да се систем нађе у одређеном стању (i -та колона - i -то стање).

Систем се у једном тренутку може налазити у само једном стању. Нека се стање система карактерише случајном променљивом X_n која у тренутку n има расподелу \vec{s} и нека је укупан број стања система M . Расподела \vec{s} је у ствари закон расподеле (енг. PMF) јер се ради о случајној променљивој дискретног типа - систем може бити само у једном од M могућих стања, и конкретно може се посматрати као вектор врсте, димензија $1 \times M$ где се на i -том месту налази вероватноћа да се систем у тренутку n нађе у стању i .

У следећем временском тренутку, $n + 1$, систем се може наћи у било ком од M стања са различитим вероватноћама. Вероватноћа да ће се систем у тренутку $n + 1$ наћи у стању j означава се са $P(X_{n+1} = j)$. Пошто ова вероватноћа зависи од стања у претходном тренутку, може се изразити на следећи начин (према формули тоталне вероватноће)

$$P(X_{n+1} = j) = \sum_i^M P(X_{n+1} = j \mid X_n = i)P(X_n = i) = *$$

$P(X_{n+1} = j \mid X_n = i)$ = вероватноћа преласка система из стања i у стање $j \rightarrow p_{i,j}$
 $P(X_n = i)$ = вероватноћа да се систем у тренутку n нађе у стању $i \rightarrow s_i$

$$* = \sum_i^M p_{i,j} s_i$$

Дакле, вероватноћа да систем у $n + 1$ -ом тренутку буде у стању j једнака је суми производа вероватноћа да се систем у n -том тренутку нађе у било ком стању и вероватноћа одговарајућих прелаза.

Ова сума представља j -ту колону у матрици (димензија $1 \times M$) која се добије при множењу вектора \vec{s} и матрице транзиције P .

Према свему наведеном следи да је закон расподеле случајне променљиве X_{n+1} (расподела вероватноћа да се систем у $n + 1$ -ом тренутку налази у сваком од стања) једнак $\vec{s} \times P$.

Аналогно, у тренутку $n+2$, случајна променљива X_{n+2} има расподелу $\vec{s} \times P^2$, у тренутку $n + 3$, случајна променљива X_{n+3} има расподелу $\vec{s} \times P^3$ итд.

Дефиниција 10 *Расподела \vec{s} за коју важи :*

$$\vec{s} \times P = \vec{s}$$

*назива се **стационарна** или **равнотежна** расподела.*

$\vec{s} \times P$ представља "један корак у будућност", тј. расподелу вероватноћа да систем нађе у сваком од стања у следећем временском тренутку. Уколико расподела остаје иста, односно, вероватноће се не мењају са временом, тада се та расподела назива стационарном. Под одређеним условима везаним Марковљеве ланце, доказује се да Марковљев ланац **увек** конвергира ка својој стационарној расподели без обзира на полазно стање. Више о конвергенцији Марковљевих ланаца може се наћи у [11]. Дакле, полазне стање се може изабрати потпуно случајно а затим, уколико се дозволи да "протекне" довољно времена, закон расподеле вероватноће да се систем нађе у свим стањима система ће конвергирати ка стационарној расподели тог ланца.

Дефиниција 11 *МСМС (енгл. Markov Chain Monte Carlo) методе представљају класу алгоритама који се користе за синтетичко генерисање узорака случајних променљивих из одговарајућих расподела. Овим методама се креирају Мерковљеви ланци који као равнотежну расподелу имају расподелу из које се узимају узорци. Једна од МСМС метода је и Гибсово узорковање (енгл. Gibbs sampling)*

Гибсово узорковање

Нека је дата заједничка расподела (енгл. joint distribution) $p(\mathbf{z}) = p(z_1, z_2, \dots, z_M)$ из које је потребно одабрати неку вредност (енгл. sample) и нека је познато почетно стање Марковљевог ланца који је потребно генерисати. Сваки корак Гибсовог узорковања почиње заменом вредности једне променљиве z_1, z_2, \dots, z_M вредношћу која се добија из **условне расподеле** те променљиве у односу на све остале. Дакле, z_i се мења вредношћу која се узима из расподеле $p(z_i \mid z_{-i})$, где је са z_i означена i -та координата вектора \mathbf{z} а са z_{-i} сви z_1, z_2, \dots, z_M без z_i . Ова процедура се наставља за све променљиве по неком одређеном редоследу. При довољном броју итерација, вредности вектора \mathbf{z} ће конвергирати ка $p(\mathbf{z})$.

На пример, нека је дата расподела три случајне променљиве $p(z_1, z_2, z_3)$ и нека су вредности у тренутку $t : z_1^t, z_2^t, z_3^t$. Нека се замена вредности променљивих врши у односу на индекс, од најмањег ка највећем. Вредност z_1^t се мења новом вредношћу z_1^{t+1} која се узима (узрокује) из расподеле

$$p(z_1 \mid z_2^t, z_3^t).$$

Сада се вредност z_2^t мења са вредношћу z_2^{t+1} која се узима из расподеле

$$p(z_2|z_1^{t+1}, z_3^t).$$

Дакле, одмах се користи нова вредност променљиве z_1 . Коначно, за промену вредности z_2^t користи се вредност z_3^{t+1} која се добија из расподеле :

$$p(z_2|z_1^{t+1}, z_3^{t+1}).$$

Овим је завршена **једна итерација** Гинбсовог узорковања. Исти процес се наставља кроз низ итерација све до одређеног броја или до неког другог услова заустављања.

Описана процедура се може уопштити и на више од три променљиве и може се представити следећим псеудокодом:

1. Initialize $\{z_i : i = 1, \dots, M\}$
2. For $\tau = 1, \dots, T$:
 - Sample $z_1^{(\tau+1)} \sim p(z_1|z_2^{(\tau)}, z_3^{(\tau)}, \dots, z_M^{(\tau)})$.
 - Sample $z_2^{(\tau+1)} \sim p(z_2|z_1^{(\tau+1)}, z_3^{(\tau)}, \dots, z_M^{(\tau)})$.
 - \vdots
 - Sample $z_j^{(\tau+1)} \sim p(z_j|z_1^{(\tau+1)}, \dots, z_{j-1}^{(\tau+1)}, z_{j+1}^{(\tau)}, \dots, z_M^{(\tau)})$.
 - \vdots
 - Sample $z_M^{(\tau+1)} \sim p(z_M|z_1^{(\tau+1)}, z_2^{(\tau+1)}, \dots, z_{M-1}^{(\tau+1)})$.

Слика 3.19: Псеудокод Гибсовог узорковања, преузето са [12]

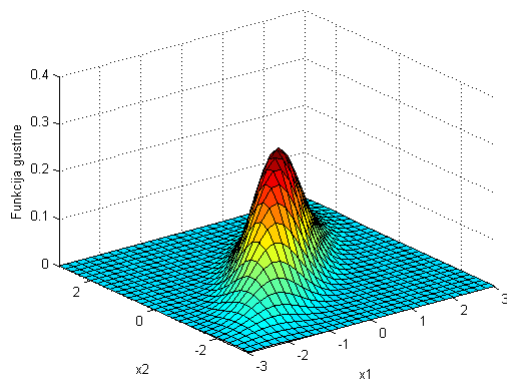
Гибсово узорковање подразумева да су унапред познате **условне расподеле** свих променљивих и да је могуће узорковање из њих.

Пример : Нека је потребно узорковати вредности из дводимензионалне нормалне расподеле $\mathcal{N}(\mu, \Sigma)$ Гибсовим узорковањем при чему је

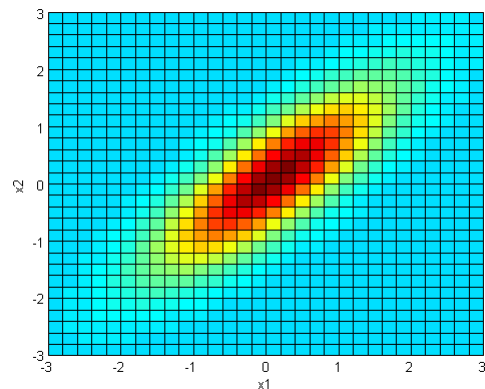
$$\mu = [\mu_1, \mu_2] = [0, 0]$$

$$\Sigma = \begin{bmatrix} 1 & \rho_{12} \\ \rho_{21} & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

Графички приказ овакве дводимензионалне нормалне расподеле дат је на следећој слици (тродимензионално и пројектовано на две димензије):



Слика 3.20: Тродимензионални приказ дводимензионалне нормалне расподеле



Слика 3.21: Дводимензионални приказ дводимензионалне нормалне расподеле

Основна претпоставка Гибсовог узорковања је да су познате условне расподеле свих променљивих и да је из њих могуће узорковати. Према [12] и [13], за условне расподеле дводимензионалне заједничке расподел важи :

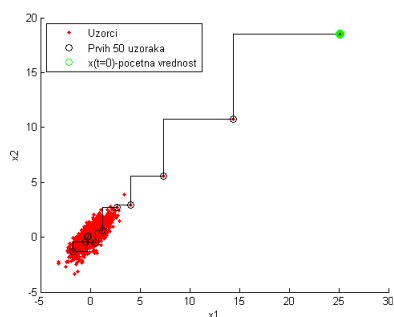
$$p(x_1 | x_2^{(t-1)}) = \mathcal{N}(\mu_1 + \rho_{21}(x_2^{(t-1)} - \mu_2), \sqrt{1 - \rho_{21}^2})$$

и

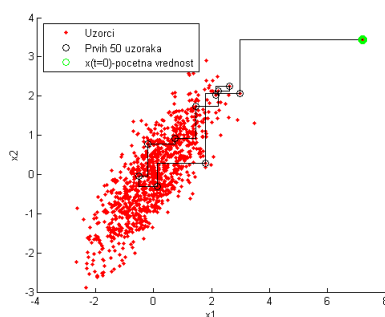
$$p(x_2 | x_1^{(t)}) = \mathcal{N}(\mu_2 + \rho_{12}(x_1^{(t)} - \mu_1), \sqrt{1 - \rho_{12}^2})$$

Дакле, обе условне расподеле представљају једнодимензионалну нормалну расподелу са одговарајућим параметрима. Почетне вредности променљивих се бирају случајно зато што нису од важности. Марковљев ланац ће свакако конвергирати ка дводимензионалној нормалној расподели са неведеним параметрима после одређеног броја итерација. У зависности од полазног стања, тај број итерација ће бити мањи или већи.

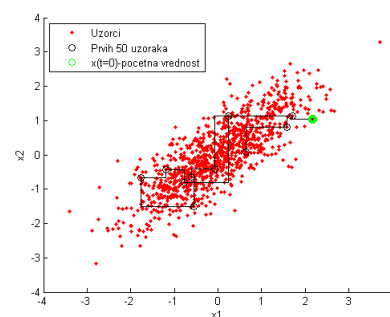
На следећем сликама су представљене добијене расподеле Гибсовим узорковањем за различите почетне вредности :



Слика 3.22: Почетна тачка (25.1284, 18.5165)



Слика 3.23: Почетна тачка (7.2162, 3.4380)



Слика 3.24: Почетна тачка (2.1649, 1.0314)

Са слика је очигледно да се првих неколико узорака може занемарити (3.22 може се занемарити првих 7-8 узорака).

Важно је приметити да се узимање узорака увек креће по "степенастом" обрасцу. Дакле, две суседне тачке имају исту једну координату (x или y). То је зато што Гибсово узорковање у једном тренутку мења **само једну** променљиву у односу на одговарајућу вредност друге.

Конвергенција алгоритма Гибсовог узорковања ка стационарној расподели Марковљевог ланца је теоретски загарантована, али је у пракси јако тешко одредити број итерација након којих ланац почиње да конвергира. Један од начина процене конвергенције је и рачунање *log-likelihood* -а

3.4 Како ради ТМ алгоритам

Раније је неформално описан LDA генеративни процес. Основна претпоставка је да се сваки документ у одређеној пропорцији говори о свакој теми (има одређену расподелу над темама) као и да свакој теми све речи из корпуса припадају са различитим вероватноћама (расподела над речима). Генеративни процес се , према [14], може описати следећим псеудокодом :

1. For $k = 1 \dots K$:
 - (a) $\phi^{(k)} \sim \text{Dirichlet}(\beta)$
2. For each document $d \in \mathbf{D}$:
 - (a) $\theta_d \sim \text{Dirichlet}(\alpha)$
 - (b) For each word $w_i \in d$:
 - i. $z_i \sim \text{Discrete}(\theta_d)$
 - ii. $w_i \sim \text{Discrete}(\phi^{(z_i)})$

Слика 3.25: Генеративни процес LDA-a

при чему је : K - укупан број тема у колекцији $\phi^{(k)}$ - расподела над свим речима из колекције и представља расподелу над речима у k -тој теми θ_d - расподела над темама у документу d . z_i - тема којој припада реч w_i . α, β - **хиперпараметри** тј. параметри симетричних Дирехлеових расподела.

Описани генеративни процес резултује формирањем следеће заједничке расподеле :

$$p(w, z, \theta, \phi \mid \alpha, \beta) = p(\phi \mid \beta) p(\theta \mid \alpha) p(z \mid \theta) p(w \mid \phi_z) \quad (3.16)$$

Непозанте променљиве које је потребно "открити" су z, θ и ϕ на основу (једино) познатих **речи** и њиховог присуства у сваком од докумената. Дакле, потребно је пронаћи расподеле наведених променљивих **под условом** да су познате речи и њихова распоређеност по документима тј. открити њихове постериорне расподеле. Основни проблем ТМ је **постериорно закључивање** (енг. posterior inference) односно отривање постериорних расподела латентних случајних променљивих на основу задатог скупа докумената и речи што представља решавање следеће једначине :

$$p(\theta, \phi, z \mid w, \alpha, \beta) = \frac{p(\theta, \phi, z, w, \mid \alpha, \beta)}{p(w \mid \alpha, \beta)} \quad (3.17)$$

Према [14], рачунање имениоца овог разломка је готово немогуће па се стога прибегава апроксимативним методама каква је и Гибсово узорковање.

Да би се применило Гибсово узорковање, потребно је познавати условне расподеле свих променљивих из чије се заједничке расподеле узоркује. Међутим, показује се да је довољно пронаћи само z јер се остале две променљиве могу преко ње израчунати и то (према [14]):

$$\theta_{d,z} = \frac{n(d, z) + \alpha}{\sum_{|Z|} n(d, z) + \alpha}$$

$$\phi_{z,w} = \frac{n(z, w) + \beta}{\sum_{|W|} n(z, w) + \beta}$$

Овако примењен алгоритам Гибсовог узорковања назива се још и енгл. Collapsed Gibbs Sampling. Дакле, циљ је пронаћи за сваку реч, вероватноћу да припадне свакој од тема, под условом да су познате теме којима припадају остале речи у том тренутку. Формалније, ово се може записати $p(z_i \mid z_{-i}, \alpha, \beta, w)$ где z_{-i} представља доделу тема свим речима сем i -те.

Априорне расподеле коришћене у ТМ су Дирихлеове. Важна особина Дирихлеове расподеле је да је она **конјугована** са мултиномијалном расподелом. Дирихлеова расподела је расподела над параметрима мултиномијалне расподеле. Нека је на почетку претпостављено да параметри мултиномијалне расподеле припадају некој Дирихлеовој расподели - $\mathbf{p} \sim \text{Dir}(\mathbf{p}, \alpha)$. Нека је \mathbf{x} узорак генерисан из мултиномијалне расподеле $\text{Mult}(\mathbf{x}; \mathbf{p})$. Тада важи да је постериорна

расподела \mathbf{p} (дакле, расподела под условом да је познат узорак \mathbf{x}) такође **Дирихлеова расподела** са параметром $\mathbf{x} + \alpha$ тј.:

$$p(\mathbf{p} \mid \mathbf{x}, \alpha) = \text{Dir}(\mathbf{p}; \mathbf{x} + \alpha) = \frac{1}{B(\mathbf{x} + \alpha)} \prod_{i=1}^{|\alpha|} p_i^{x_i + \alpha_i - 1} \quad (3.18)$$

Како (3.18) представља **расподелу** то важи да је :

$$1 = \int \frac{1}{B(\mathbf{x} + \alpha)} \prod_{i=1}^{|\alpha|} p_i^{x_i + \alpha_i - 1} = \frac{1}{B(\mathbf{x} + \alpha)} \int \prod_{i=1}^{|\alpha|} p_i^{x_i + \alpha_i - 1} \quad (3.19)$$

Одакле следи да је :

$$\int \prod_{i=1}^{|\alpha|} p_i^{x_i + \alpha_i - 1} = B(\mathbf{x} + \alpha) \quad (3.20)$$

Ова једнакост је важна за даљи опис рада ТМ алгоритма.

Према формули условне расподеле, важи :

$$p(z_i \mid z_{-i}, \alpha, \beta, w) = \frac{p(z_i, z_{-i}, w \mid \alpha, \beta)}{z_{-i}, w \mid \alpha, \beta)} \propto p(z_i, z_{-i}, w \mid \alpha, \beta) = p(z, w \mid \alpha, \beta) \quad (3.21)$$

$p(z, w \mid \alpha, \beta)$ се може посматрати као "маргиналан расподела" две променљиве заједничке расподеле (3.16) па важи :

$$p(z, w \mid \alpha, \beta) = \iint p(z, w, \theta, \phi \mid \alpha, \beta) d\theta d\phi = \iint p(\phi \mid \beta) p(\theta \mid \alpha) p(z \mid \theta) p(w \mid \phi_z) d\theta d\phi \quad (3.22)$$

Груписањем по заједночкој зависној променљивој, претхонда једначина се може написати :

$$p(z, w \mid \alpha, \beta) = \int p(z \mid \theta) p(\theta \mid \alpha) d\theta \int p(w \mid \phi_z) p(\phi \mid \beta) d\phi \quad (3.23)$$

Оба интерграла представљају комбинацију узорка из мултиномијалне расподеле и априорне Дирихлеове расподеле. Како је Дирихлеова расподела конјугована (conjugate prior) са мултиномијалном, у подинтегралном изразу се налази "множење" две Дирихлеове расподеле са одговарајућим параметрима.

Дакле : Пошто $p(z \mid \theta)$ има мултиномијалну дистрибуцију, важи:

$$p(z \mid \theta) = \prod_{i=1}^D \prod_{k=1}^K \theta_{d,k}^{\Omega_{d,k}} \quad (3.24)$$

, где је $\Omega_{d,k}$ број који означава колико пута је тема k додељена у документу d - број речи који у документу d припадају теми k .

Члан $p(\theta \mid \alpha)$ је из основне Дирихлеове расподеле па важи :

$$p(\theta \mid \alpha) \stackrel{(1)}{=} \prod_{i=1}^D p(\overline{\mathbf{q}}_d \mid \alpha) \stackrel{(2)}{=} \prod_{d=1}^D \frac{1}{B(\alpha)} \prod_{k=1}^K q_{d,k}^{\alpha_k - 1} \quad (3.25)$$

где је $\overline{\mathbf{q}}_d$ расподела вероватноћа тема у документу d . Расподеле вероватноћа тема по документима су независне, па је зато могуће написати (1). Расподела тема по документу се узима из Дирихлеове расподеле па је зато могуће написати (2).

Према томе, први интеграл једнакости (3.23) се записује као :

$$\int p(z | \theta) p(\theta | \alpha) d\theta = \int \prod_{i=1}^D \prod_{k=1}^K \theta_{d,k}^{\Omega_{d,k}} \prod_{d=1}^D \frac{1}{B(\alpha)} \prod_{k=1}^K q_{d,k}^{\alpha_k-1} d\theta_d \stackrel{(1)}{=} \prod_{i=1}^D \int \frac{1}{B(\alpha)} \prod_{k=1}^K q_{d,k}^{\Omega_{d,k}+\alpha_k-1} d\theta_d \quad (3.26)$$

Једнакост (1) следи из чињенице да су θ_d независне расподеле па се могу интегралити посебно - правило интеграције производа Према релацији (3.20) претходна једнакост се може написати и као се :

$$\int p(z | \theta) p(\theta | \alpha) d\theta = \prod_{i=1}^D \frac{B(\Omega_d + \alpha)}{B(\alpha)} \quad (3.27)$$

где је са Ω означена матрица докумената и тема, $\Omega_{d,k}$ означава колико је пута тема k додељена у документу d а Ω_d је d -та врста те матрице. Елементи ове матрице могу се математички записати и овако :

$$\Omega_{d,k} = \sum_{i=1}^N I(d_i = d \wedge z_i = k) \quad (3.28)$$

где је N укупан број речи у корпусу (са понављањем).

Аналогно претходним извођењима, и други интеграл може да се упрости:

Члан $p(\phi | \beta)$ је из основе Дирхлеове расподеле па важи :

$$p(\phi | \beta) = \prod_{k=1}^K p(\phi_k | \beta) = \prod_{k=1}^K \frac{1}{B(\beta)} \prod_{v=1}^V \phi_{k,v}^{\beta_v-1} \quad (3.29)$$

Члан $p(w | \phi_z)$ има мултиномијалну расподелу па важи :

$$p(w | \phi_z) = \prod_{i=1}^N p(\phi_{z_i, w_i}) = \prod_{k=1}^K \prod_{v=1}^V \phi_{k,v}^{\Psi_{k,v}} \quad (3.30)$$

где је ψ $K \times V$ матрица а $\psi_{k,v}$ броји колико тема k била додељена речи v . Ова матрице може се још написати као :

$$\psi_{k,v} = \sum_{i=1}^N I(w_i = v \wedge z_i = k) \quad (3.31)$$

На основу (3.29) и (3.30) следи :

$$\int p(w | \phi_z) p(\phi | \beta) d\phi = \int \prod_{k=1}^K \frac{1}{B(\beta)} \prod_{v=1}^V \phi_{k,v}^{\psi_{k,v}+\beta_v-1} d\phi_k \quad (3.32)$$

Аналогно извођењу (3.26) (3.27) следи :

$$\int \prod_{k=1}^K \frac{1}{B(\beta)} \prod_{v=1}^V \phi_{k,v}^{\psi_{k,v}+\beta_v-1} d\phi_k = \prod_{k=1}^K \left(\frac{1}{B(\beta)} \int \prod_{v=1}^V \phi_{k,v}^{\psi_{k,v}+\beta_v-1} d\phi_k \right) = \prod_{k=1}^K \frac{B(\psi_k + \beta)}{B(\beta)} \quad (3.33)$$

Коришћењем једнакости (3.27) и (3.33), једнакост (3.23) се може записати као :

$$p(z, w | \alpha, \beta) = \prod_{i=1}^D \frac{B(\Omega_d + \alpha)}{B(\alpha)} \prod_{k=1}^K \frac{B(\psi_k + \beta)}{B(\beta)} \quad (3.34)$$

На основу (3.34) може се извести правило по коме ће Гибсов алгоритам узорковања мењати доделе тема речима. Дакле :

$$p(z_i = k | Z^{-i}, W, \alpha, \beta) = \frac{p(z_i = k, Z^{-i}, W | \alpha, \beta)}{p(Z^{-i}, W | \alpha, \beta)} = \frac{p(Z, W | \alpha, \beta)}{p(Z^{-i}, W | \alpha, \beta)} \quad (3.35)$$

Именилац претходне једнакост се може написати преко условне вероватноће у следећем облику :

$$p(Z^{-i}, W | \alpha, \beta) = p(Z^{-i} | \alpha\beta)p(W | Z^{-i}, \alpha\beta) \stackrel{(1)}{=} \quad (3.36)$$

$$= p(Z^{-i})p(W^{-i} | Z^{-i})p(w_i) \propto p(Z^{-i})p(W^{-i} | Z^{-i}) = p(Z^{-i}, W^{-i}) \quad (3.37)$$

Једнакост (1) следи из чињенице да свако z_i зависи само од w_i . Од ове једнакости параметри α, β су изостављени због прегледности, али се подразумевају.

Форма једнакости (3.37) иста је као (3.31) па се једнакост (3.35) записује као :

$$p(z_i = k | Z^{-i}, W, \alpha, \beta) = \prod_{k=1}^K \frac{B(\psi_k + \beta)}{B(\psi_k^{-i} + \beta)} \prod_{d=1}^D \frac{B(\Omega_d + \alpha)}{B(\Omega_d^{-i} + \alpha)} \quad (3.38)$$

Коришћењем особина бета фунцкије, претходна једнакост се своди на :

$$p(z_i = k | Z^{-i}, w = v, W^{-i}, \alpha, \beta) = \frac{\psi_{k,v} + \beta_{w_i} - 1}{\left[\sum_{v=1}^V \psi_{k,v} + \beta_v \right] - 1} [\Omega_{d,k} + \alpha_k - 1] \quad (3.39)$$

Детаљно извођење може се наћи код [14] и [15].

3.4.1 Имплементација - псеудокод

Овде ће доћи мој псеудокод, ово је само привремено

```

Input: words  $\mathbf{w} \in$  documents  $\mathbf{d}$ 
Output: topic assignments  $\mathbf{z}$  and counts  $n_{d,k}, n_{k,w}$ , and  $n_k$ 
begin
  randomly initialize  $\mathbf{z}$  and increment counters
  foreach iteration do
    for  $i = 0 \rightarrow N - 1$  do
       $word \leftarrow w[i]$ 
       $topic \leftarrow z[i]$ 
       $n_{d,topic} -= 1; n_{word,topic} -= 1; n_{topic} -= 1$ 
      for  $k = 0 \rightarrow K - 1$  do
         $p(z = k | \cdot) = (n_{d,k} + \alpha_k) \frac{n_{k,w} + \beta_w}{n_k + \beta \times W}$ 
      end
       $topic \leftarrow \text{sample from } p(z | \cdot)$ 
       $z[i] \leftarrow topic$ 
       $n_{d,topic} += 1; n_{word,topic} += 1; n_{topic} += 1$ 
    end
  end
  return  $\mathbf{z}, n_{d,k}, n_{k,w}, n_k$ 
end

```

Algorithm 1: LDA Gibbs Sampling

Слика 3.26: Псеудокод - преузето са [15]

Глава 4

Историја ТМ

Глава 5

Опис приступа решавању проблема

5.1 Припрема подарака - предпроцесирање

Подаци кориштени за тестирање приликом израде овог рада су јавно доступни подаци са сајта <http://stackexchange.com/> из три области - инжињерство, фитнес и хемија. Подаци су на **енглеском језику**.

Ове три области су намерно тако одабране како би се повећала разноврсност речи. Речи из сродних научних грана користе сличну или исту терминологију па опасност од бирања питања и одговора из сродних научних дисциплина лежи у чињеници да ће диверзитет корпуса бити мали.

Из сваке области узето је по 200 питања и одговора, што представља базу од укупно 600 питања и 600 одговора. Овакав скуп података је затим подељен на два дела - тренинг део (360 питања и одговора) и тест део (240 питања и одговора).

Подаци добијени на овај начин нису погодни за директну обраду те их је потребно делимично **прерадити** тј. **предпроцесирати**.

Приликом решавања проблема коришћен је пакет **Mallet** и његова решења за алгоритам моделовања тема, предпроцесирање података итд. Више о овом пакету може се пронаћи у [16]

5.1.1 Уклањање HTML ознака и неалфанумеричких карактера

Подаци који су доступни преко сајта <http://stackexchange.com/> дати су у форми HTML текстова. То значи да се кроз основни текст питања и одговора провлаче додатни низови карактера који представљају HTML ознаке (тагове). Ове ознаке добијају смисао приликом генерисања веб странице, односно приликом форматирања текста питања и одговора на веб страницама. Међутим, за обраду текста у овом раду, те ознаке не значе ништа и додатно могу уносити забуну. Због тога их је потребно, пре било какве даље обраде текста, склонити. Најједноставнији начин да се то уради је преко *регуларних израза*, обзиром на изглед и формат HTML ознака.

Са друге стране, у тексту се поред слова и бројева (*алфанумерички карактери*) обавезно појављују и специјални знаци (ознаке које нису ни слова ни бројеви - *неалфанумерички карактери*, међу којима су најчешћи интерпункцијски знакови). Поред њих, у скорије време, уз сваки текст је готово обавана појава и група специјалних знакова као што су нпр. смајлићи (структуре конструисане од специјалних знакова које се могу интерпретирати као расположење нпр. :) - срећа, :(- туга итд.). Са становишта алгоритма моделовања тема, сви они, исто као и HTML ознаке, немају значење и потребно их је уклонити. Такође, најједноставнији начин да се то уради је применом регуларних израза.

5.1.2 Конвертовање свих слова текста у „мала слова“ енг. lowercase

Писање речи великим или малим словима, као и започињање речи малим или великим словом, углавном има граматички смисао. При томе, у већини случајева, реч написана почетним великим и иста реч написана почетним малим словом, имају исто значење. У машинској обради података, свако слово има другачију бинарну репрезентацију. Стога, реч написана великим почетним словом и иста реч написана малим почетним словом или свим великим словима, **нису исте речи** обзиром на то да имају различиту репрезентацију. Како би се ово избегло, пре било какве даље обраде, врши се конвертовање или пребацивање свих слова текста у мала слова. На тај начин, све речи састављене од истих слова и у истом редоследу, представљају **исту** реч, без обзира на граматичко значење речи или њену позицију у тексту. Избор конвертовања у мала слова је једнако оправдан као и конвертовање у велика слова. Дакле, исти резултат би се добио и конвертовањем свих слова у велика слова. Међутим, у пракси, је чешће прихваћено пребацивање у мала слова па је такав приступ усвојен и у овом раду. Наравно, постоје бројни примери када писање речи великим или малим почетним словом битно мења значење речи. На пример, у српском језику, реч *Мила*, написана великим почетним словом, означава име особе, именуцу, док реч *мила*, написана малим почетним словом означава придев. Исто тако реч, *Јела* односи се на име особе док се реч *јела* односи на врсту зимзеленог дрвета. Међутим, оваквих речи има довољно мало да је ризик од мењања значења речи, са аспекта алгорита моделовања тема, прихватљив.

5.1.3 Издавање атомских елемената докумената - токена, енг. tokenization

Манипулација читавим документима са аспекта алгорита моделовања тема, нема смисла. Основна јединица манипулације ове врста алгоритама је **реч**. Према томе, потребно је документе рашчланити на појединачне речи. Процес издавања основних елемената манипулације, тј. елемената од интереса, назива се издавање **токена** или **токенизација**. Обзиром да су овде атомски елементи **речи** и да су речи одвојене размаком, процес токенизације је најједноставније извршити преко регуларних израза. У пакету Mallet се већ налазе готове класе

5.1.4 Избацивање често коришћених речи енг. stop words

У свакодневном говору често се употребљавају личне заменице, прилози, везници итд. Без њих, говор би био неодређен и неповезан па самим и неразумљив. Међутим, у машинској обради текстуланих података, поготово у алгоритмима моделовања тема, оваква врста информација није неопходна. Пре свега, такве речи не носи суштинско теметско значење обзиром да нису уско повезана ни са једном конкретном области. На пример, везници ,као што су у српском језку: и, или, па, али, због, ради итд. се употребљавају при писању докумената из свих научних грана и стога се ни за једну од тих речи не може дефинисати област припадања. Обзиром да је циљ из групе докумената издвигнути **теме**, овакве речи су сувишне. Штавише, уносе додатну забуну при закључивању и, обзиром да су бројене, могу представљати велико оптерећење приликом обраде. Подаци коришћени у при изради овог рада су на енглеском језику те ће се надаље говорити о оваквој врсти речи у енглеском језику. Међутим, узимањем уобзор специфичности конкретног језика, иста разматрања могу се применити и на друге језике.

Избацивање често коришћених речи, stop words-ова, може се реализовати на више начина.

Поузданији, универзалнији али и рачински захтевнији начин је алгоритаМСКО проналажење таквих речи. Обзиром да немају тематско значење, оне се појављују у великом броју у свим документима и темама. Једноставним бројањем појављивања речи у скупу свих докумената могу се уочити групе речи које се са изузетно високим фреквенцијама јављају

у **свим документима**. Такве речи се могу сматрати за често коришћене те се, из описаних разлога, избацују. Поред једноставног бројања речи, постоје и друге методе „ мерења “ присуства речи у корпусу. Једна од њих је и релативна фреквенција која зависи од дужине докумената. Међутим, обзором да овај приступ није коришћен у раду, у ове методе се неће дубље улазити.

Други, мање поуздан и релативно рачунски незахтеван приступ је коришћење **листе често коришћених речи** које постоје за сваки језик. Те листе су јавно доступне и могу се пронаћи на бројним веб сајтовима. Обзором да је циљ рада био истраживање примене алгоритама моделовања тема у специфичном проблему, овај приступ је прихватљивији. Пре свега, релативно се лако имплементира обзором да у Mallet-у већ постоји класа за уклањање ових речи. Са друге стране, овај корак предпроцесирања се на овај начин претвара у тривијалан и оставља простор за истраживање самог алгорита моделовања тема. Највећа опасности од овог приступа је елиминација речи које, иако сврстане међу често коришћене, у датом скупу докумената ипак имају значење. Исто тако, обзиром да је листа предефинисана, могуће је изоставити речи које у конкретном скупу представљају често коришћене речи.

За поузданије и детаљнија истраживања, предлаже се примена прве методе.

5.1.5 Додавање синонима

У циљу бољег дифренцирања тематике питања и одговора, за сваку реч је додато по 5 синонима. За проналажење синонима је коришћена WordNet библиотека. Основни разлог додавања синонима у скуп била је претпоставка да ће се на тај начин боље диференцирати теме, повећати диверзитет корпуса а самим тим и олакшати препознавање тачног одговора. Међутим, резултати су показали управо супротно. Разлог томе што синониме треба тражити **по смислу** речи а не само по лексичком облику речи. Такође, фразе, којих има доста у свакодневном говорном и писаном енглеском језику, значајно могу да утичу на смисао питања/одговора. Када се они рашчлане на појединачне речи, могуће је да се и смисао промени.

5.1.6 Склањање наставака речи - енг. *stemming*

Овај сегмент предпроцесирања је изузетно завистан од језика на коме је текст писан. Циљ је препознати различите облике исте речи и свести је на заједничку основу, која **не мора** да буде коренска. У енглеском језику, различити облици речи граде се додавањем разних **наставака** као што су *s,ing,es* Дакле, склањањем наставака речи редукује се диверзитет корпуса али се истовремено речи које имају исто значење само различит облик настао услед контекста реченице - рода, времена, врсте речи итд, свде на исту реч. У конкретној примени, овако нешто је неопходно за прецизно раздвајање тема. За српски језик овако нешто не би било могуће имплементирати на једноставан начин обзором на промену речи по падежима, лицима (за глаголе), родовима, бројевима и гласовне промене које се при томе дешавају. У тренутку писања рада, никакво готово решења за српски језик није постојало. Уколико предмет истраживања неког будућег рада буде био српски језик, потребно је написати процедуре којима се речи језика ослобађају наставака, следећи граматичка правила.

5.1.7 Свођење на коренску реч - енг. *lemmitization*

Свођење речи на коренску реч је слична метода методи склањања наставака речи, такође уско повезана са језиком који се обрађује. Једна од разлика је што се склањање наставака речи може применити на речи које мењају облик додавањем наставака док се свођење на коренски реч може применити на све речи. На пример, реч енг. *better* - бољи, при

склањању наставака би остала непромењена или би се свела на реч енг. *bett*, док при свођењу на коренску реч она постаје енг. *good* - добар. Друга, битнија разлика, је што се уклањање наставака примењује на речи не водећи рачуна о контексту, док се свођењем на коренску реч може специфицирати и контекст речи. На пример, реч енг. *meeting* - може имати више значења. Као именица она означава *састанак* док као глагол означава презент партицип глагола *to meet*, у смислу *сретати се*, . Уклањањем наставака, реч енг. *meeting* у контексту именице као и у контексту глагола биће сведена на реч енг. *meet*. Код свођења речи на коренску реч, спецификацијом врсте речи могуће је реч енг. *meeting* оставити непромењену. Такође, свођењем на коренску реч могуће је неправилне глаголе енглеској језика свести на основни облик, што склањањем наставака није било могуће. Са аспекта алгоритма моделовања тема, свођење на коренску реч је прихватљивија метода, пре свега због могућности препознавања различитог облика истих речи када се они не граде додавањем наставака. Обзором на једноставност имплементације и брзину рада, чешће се користи скалањање наставка од свођења на коренску реч. У конкретном раду, обе методе су независно тестиране и равноправно коришћене у циљу добијања бољих резултата.

5.2 Развој решења

5.2.1 Општи преглед пакета Mallet

5.2.2 Pipe класа Mallet-у

5.2.3 ParallelTopicModel класа Mallet-у

5.3 Тестирање решења

5.3.1 Општи приступ тестирању

5.3.2 Начин тестирања у конкретном проблему

Глава 6

Решење проблема применом методе бројања речи

Метода бројања речи једна је од једноставнијих метода којом је могуће поредити два текстулана документа. Основна претпоставка ове методе је да су документи ближи један другом уколико имају више заједничких речи. Оваква претпоставка иако делује сасвим основано, не мора увек да буде тачна. Истим скупом речи могу се описати потпуно различите ствари и тиме генерисати два текстулана документа која, по смислу, уопште нису слична. Иако су овакви примери бројни, ова метода је, пре свега, због једноставности имплементације широко прихваћена у системима за проналажења одговора на постављено питање.

У конкретном раду, метода бројања речи је коришћена као **компаративно решење** у односу на решење применом алгоритма моделовања тема.

6.1 Опис решења методом бројања речи

Улаз у алгоритам су група питања и група одговора, при чему се за свако питање унапред зна који одговор из дата групе одговора представља тачан одговор. Решење методом бројања речи заснива се на мерењу сличности датог питања са **сваким одговором** у бази одговора. Након тога, одговори се рангирају према израчунатој сличности. Позиција тачног одговора у тој хијерархији свих одговора представља излаз који алгоритам даје за свако постављено питање.

Глава 7

Резултати

7.0.1 Утицај стеминга на резултат

7.0.2 Утицај додавања синонима на резултат

7.0.3 Укупни резултати са синонимима и стемингом

Глава 8

Коришћене скраћенице

1. енг. - енглески језик
2. срп. - српски језик
3. LDA - Latent Dirichlet Allocation
4. ТМ - Topic Model, срп. моделовање тема

Литература

- [1] Lars Vogel, Android Service and Broadcast Receiver, www.vogella.de, 2011.
- [2] David M. Blei, Introduction to Probabilistic Topic Models, Princeton University
- [3] David M. Blei, Topic Models, Princeton University, September 1, 2009
- [4] David Mimno, The details: training and validating big models on big data, Princeton University
- [5] Ivana Kovačević, Verovatnoća i statistika sa zbirkom zadataka, Beograd 2011.
- [6] Violeta Aleksić, Elementi teorije verovatnoće i matematičke statistike,
- [7] www.ekfak.kg.ac.rs, kurs Osnovi statistike, avgust 2015.
- [8] <http://starisajt.elfak.ni.ac.rs/phptest/new/html/Studije/predavanja-literatura/matematika-odabrana-poglavlja/verovatnoca.pdf>, avgust 2015
- [9] Random Signals and Processes Primer with MATLAB, Gordana Jovanovic Dolocek, 2013
- [10] Lecture 32: Markov Chains Continued | Statistics 110 on youtube - <https://www.youtube.com/watch?v=aBGOyZv2pZE>, avgust 2015
- [11] Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006
- [12] <http://fourier.eng.hmc.edu/e161/lectures/gaussianprocess/node7.html>, avgust 2015
- [13] <https://theclevermachine.wordpress.com/2012/11/05/mcmc-the-gibbs-sampler/>, avgust 2015
- [14] William M. Darling, A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling, December 1, 2011
- [15] Yi Wang, Distributed Gibbs Sampling of Latent Topic Models: The Gritty Details, August , 2008
- [16] <http://mallet.cs.umass.edu> , Avgust 2015

УНИВЕРЗИТЕТ У КРАГУЈЕВЦУ
ПРИРОДНО-МАТЕМАТИЧКИ ФАКУЛТЕТ
ИНСТИТУТ ЗА МАТЕМАТИКУ И ИНФОРМАТИКУ

Завршни рад под називом
одбраћен је _____.

МЕНТОР:

др Име Презиме, звање, Институција

ЧЛАНОВИ КОМИСИЈЕ:

др Име Презиме, звање, Институција

др Име Презиме, звање, Институција

Завршни рад је оцењен оценом _____.